CORPUS LINGUISTICS AND THE AUTOMATIC ANALYSIS OF ENGLISH

NELLEKE OOSTDIJK

CORPUS LINGUISTICS AND THE AUTOMATIC ANALYSIS OF ENGLISH

Printed in The Netherlands.

CORPUS LINGUISTICS AND THE AUTOMATIC ANALYSIS OF ENGLISH

een wetenschappelijke proeve op het gebied van de letteren, in het bijzonder de taalwetenschap

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Katholieke Universiteit te Nijmegen, volgens besluit van het college van decanen in het openbaar te verdedigen op woensdag 2 oktober 1991 des namiddags te 1.30 precies

door

Nelly Hendrika Johanna Oostdijk

geboren op 24 september 1958 te Nijmegen

Rodopi

Amsterdam - Atlanta, GA 1991

Promotor: Prof. dr. J.M.G.A. Aarts

for my parents

Contents

Acknowledgements

I would like to thank all those who have contributed to this book in one way or another. Special thanks are due to

Jan Aarts, supervisor of the Nijmegen corpus linguistics research group, for his guidance and encouragement;

my collaborator in the TOSCA projects, Theo van den Heuvel, and also the student assistants we have worked with over the years, for their assistance;

Joop Leo, Hans Meijer and Bas Peeters, who showed their patience in introducing me to the world of computer science and who never failed to help me out;

Hans van Halteren for his attempts to make things run more smoothly;

Flor Aarts, Pieter de Haan, Jos Hallebeek, and Gerard Kempen for reading an earlier draft of this book and making valuable comments and suggestions.

Finally, I would like to thank Peter Beinema whose unfailing support was perhaps too often taken for granted.

"... linguistics is perhaps most itself and preserves its integrity when it focuses on language as human, social behaviour, when it has a balanced concern both with modelling linguistic competence and with what actually happens in situations, patterns discoverable in the records of language events. A constant recourse to the records of such events, records of both phonic and graphic substance and of possibly relevant extratextual features, is perhaps the key to balance; the language event being both a manifestation of competence and an instance of performance must remain our point of departure and return."

-- Gregory, 1967: 197 --

Preface

In recent years, corpus linguistics -- the branch of linguistics that is concerned with the study of language use by means of large text corpora -- has developed into a discipline in its own right, while continuing to be an important ancillary discipline to various other linguistic subdisciplines. Having emerged from the convergence of computational linguistics and descriptive linguistics, corpus linguistics stands out from other branches in linguistics mainly because of its methodology. It has gained access to linguistic data that previously could only be obtained on a very small scale or not at all while at the same time various techniques for the manipulation and interpretation of data have been discovered and/or further developed.

This book is an account of some aspects of the research that was carried out at Nijmegen University from 1981 to 1989. While during this time several research projects in the field of corpus linguistics were embarked upon, including projects aimed at the analysis of Modern Standard Arabic and European Spanish, the account presented here is restricted (basically) to the two TOSCA projects in which the author herself took part. The first TOSCA project was a four-year project, funded by the University of Nijmegen Research Fund (UOP L2/80), which started in 1981. It aimed at the design and implementation of computational tools for the automatic syntactic analysis of corpora. Prior to that, there had been very little experience with the large-scale analysis of corpora, both in Holland and abroad. Whatever experience there was, was mainly with the lexical and morphological analysis of corpora. The Dutch Computer Corpus Pilot Project, which the Department of English of Nijmegen University had initiated and taken part in prior to the TOSCA project, was aimed at the syntactic analysis of a 130,000 word corpus of English. In the course of the project the need for computational tools geared to the large-scale syntactic analysis of corpora had become apparent. The TOSCA project therefore was undertaken to develop an interactive system that would enable linguists "to process large untagged corpora of texts in such a way that they will produce and retain detailed syntactic information" (Aarts and van den Heuvel, 1982: 73). Since a parser was to constitute an integral part of the system, a second aim consisted in providing a formalism suitable for writing linguistically motivated grammars that could be automatically converted into parsers. Upon completion of this project, the TOSCA II project was started. Funded by the Dutch Research Council for Advanced Research (NWO grant no. 300-169-005), it aimed at the syntactic analysis of a one million word-corpus of contemporary English. In the analysis the tools were employed that had been developed in

the preceding project.

The first chapter introduces the subject of corpus linguistics. It discusses the nature of this interdisciplinary approach to linguistics, including its goals and methodology. While attention is given to aspects of the common interest in natural language processing held by corpus linguistics and other branches of computational linguistics alike, differences that occur between these approaches are pointed out. Moreover, a comparison between two approaches in corpus linguistics, the one probabilistic, the other non-probabilistic, serves to further identify the nature of corpus linguistics 'Nijmegen-style'.

Corpus linguistics at Nijmegen University, as it has developed over the years, distinguishes itself from other corpus-based approaches mainly in that it is essentially descriptive linguistics; its principal incentive is a keen interest in language use and language variation. The tools that were developed, among which the analysis system referred to above, grant access to a wealth of information that may help to deepen our insights in this matter. A central role in this approach is played by two of the main input components to the system, i.e. the corpus and the grammar. These matters are gone into in the first chapter. Chapter two describes what criteria were involved in the compilation of the Nijmegen TOSCA Corpus, a 1.5 million word-corpus intended for studying linguistic variation. The grammar employed in the analysis of the material forms the topic of the third and the fourth chapters. The third chapter discusses some of the design issues that play a role in developing a grammar to be used in a corpus linguistic setting. Following a more general discussion of the role and the nature of such a grammar, the objectives aimed for and requirements made, a more specific introduction is given to the formalism of Extended Affix Grammar and the structure of the grammar as it was used in the analysis. Next, chapter four focuses on some issues in the implementation of the grammar by means of a discussion of the description of coordination and gapping, and the noun phrase. Finally, chapter five evaluates a number of aspects that relate to the functioning of the grammar and may be considered in future research.

1 Introduction

1.1 Natural language processing

Over the past decades, as computers became available on a very large scale indeed, while computer time, speed and memory space did no longer constitute a serious hindrance to more practical applications, computers came to occupy a place in society around which many activities today revolve. As the computer became a common commodity, the call for stepping up its performance, not only in terms of computing power but also in terms of 'intelligence' prevailed. The design of higher order programming languages, database query systems, etc. can be viewed as attempts at, on the one hand, optimizing data storage and handling, and, on the other hand, improving on the ease with which this can be done. Amid these developments the interest in various forms of man-machine interaction, from straightforward question-answering systems to highly advanced natural language interfaces, was taken up by (among others) the discipline of computational linguistics.

In the 1970s and 1980s the interest in natural language processing (NLP) has been booming. During this time a remarkable change in approach came about. Whereas at first most research in this field was hardly linguistically oriented, it was gradually realized that language itself might hold the key to success. The incorporation of essentially linguistic components, such as grammars, morphological transducers, lexicons, etc. was the answer to the failure of the earlier overall systems in which, for example, as Gazdar (1985: 186) observes, "the parser and the grammar would be thoroughly intermingled in monolithic hunks of code that have proved impossible to maintain."

The strand of computational linguistics described above is perhaps best characterized as application-oriented, since it aims at the design and implementation of computational tools for the benefit of automation. The work that has been done in this area has resulted in a large variety of systems that were developed for specific applications. Consequently, these systems are generally found to be domain-limited. For an overview we refer to Winograd (1983: 357-410).[1] Two other strands of computational linguistics, however, must be mentioned here.

---

[1] Winograd presents an overview of some 50 odd computer systems that were developed in such areas as machine translation, data base retrieval, text analysis and/or generation, etc.

One consists in using the computer as a tool in the testing of theoretical linguistic models. To the domain of this particular branch of the discipline belongs the development of theory-motivated systems. Thus over the years program suites were developed for testing transformational grammars (Friedman, 1969), Montague grammars (Friedman, 1978), Generalized Phrase Structure grammars (Evans, 1985; Phillips and Thompson, 1985), etc. The other strand is best characterized as corpus-based computational linguistics. Contrary to the essentially theoretical approach, as well as the other computational linguistic approaches, corpus linguistics distinguishes itself from these other approaches in that it takes an interest in language itself as it is actually produced, its structure, its use and variation in that use. As such it can be looked upon as a continuation of the tradition of descriptive linguistics as reflected, for example, in The Great Tradition.[2] It is computational linguistics in that it employs the computer in order to investigate large bodies of language material, so-called corpora. Its primary aim, however, lies in providing an adequate description of the corpus language. Computational techniques are thus merely a means to an end.[3]

1.2 Corpus linguistics

By defining corpus linguistics as a branch of computational linguistics it becomes immediately apparent that it has little to do with earlier approaches in linguistics that made use of corpus data. Many of these suffered from the fact that the analysis of the data was performed by hand. Descriptions underlying the analysis lacked any formal basis, while the corpora that were used were commonly rather small, privately owned collections of data, accessible only to few people. Consequently, much of the research that was done then proves impossible to verify, is inconsistent, and because of the fact that it was carried out on a relatively small scale, it is hardly conclusive about anything.

---

[2] The traditional comprehensive grammars of the English language that were written in the first half of the twentieth century are collectively referred to as 'The Great Tradition'. Among these are the grammars by Kruisinga, Poutsma, Jespersen, and Quirk et al.

[3] With respect to the aims pursued in corpus linguistics there exist rather divergent views. The view presented here is the one we hold at Nijmegen. Others are discussed below.

Although we said above that computational techniques were merely a means to an end, the use of the computer has a very large impact on the methodology that has been adopted. For one thing, at least in the view that we hold at Nijmegen, corpus linguistics today must be characterized as a formalized approach to descriptive linguistics. In its raw form, the computer-readable corpus serves as a test-bed for the linguistic hypotheses that are laid down in a formal grammar. Once analyzed, the corpus constitutes a database that may be consulted in order to obtain information about linguistic structures, their frequency and distribution, as well as to gain insights into the co-occurrence restrictions that hold. Note that this approach differs considerably from others, such as the one adhered to by, for instance, linguists who work at or in collaboration with Lancaster University's UCREL.[4] Within the corpus-based paradigm in computational linguistics theirs is a probabilistic approach. This means that at the basis of the automatic language-processing system they are developing one does not find a generative grammar, describing the set of potential sentences in a particular language[5]; rather, a 'constituent-likelihood grammar' is used, where the likelihood of a particular analysis for a given utterance is derived from empirical statistics concerning the (observed) relative frequency of occurrence of particular structures. The strength of this approach, according to Leech (1987: 3), "is that, through probabilistic predictions, it is able to deal with any kind of English language text which is presented to it: it is eminently robust. Its weakness is that the very reliance on probability admits the possibility of error. The probabilistic system makes the best 'guess' available to it, based on textual material that has been analysed in the past." The Lancaster choice for a probabilistic approach can best be explained through the long term research goals the Lancaster group has set itself, which lie within the domain of man-machine interface research "where the goal is to produce computer systems which will accept any input in a given natural language" (Leech, 1987: 4). The Lancaster approach then is primarily aimed at developing or supporting computer applications. In this respect it is much closer to other strands of computational research than to the corpus-based approach followed at Nijmegen.[6]

[4] The acronym stands for the Unit for Computer Research on the English Language.

[5] A further discussion of these two corpus-based approaches may be found in chapter 3.

1.3 The data

As was observed in section 1.1 the main objective of corpus linguistics today is the study of actual language use and hence, of language variation. The corpus, a collection of stretches of connected discourse in a single dialect, constitutes the principal source of data. The corpus is a record of performance: the utterances contained in it are unsolicited historical linguistic events and as such to be distinguished from other data, such as potential utterances or utterances that originate from experiments in a laboratory environment. While the corpus is the main source of data available to the corpus linguist, it is not the only source: an important role is assigned to his intuitions about the language. In current corpus linguistic practice, only part of the linguist's intuitions are contained in the formal grammar that is used for analysis. It is only his intuitions about the syntax that are given formal expression in the grammar -- although these are, of course, based on and integrated in his total knowledge of the language. The intuitions concerning the semantic and/or pragmatic interpretation of utterances have not (yet) been incorporated in our formal grammar and can therefore only be brought into play by way of interventions.

Unlike earlier corpora, the corpora that are currently used are computer readable and lend themselves to automatic analysis. As a result, larger quantities of data can be processed at greater speed, while consistency in the analysis is warranted through the use of a formalized description contained in the grammar.

---

[6] Sampson, contrasting it with other approaches in computational linguistics, characterizes the 'Lancaster approach' as follows:

> "The hallmarks of our approach are: (i) analytic techniques which depend on statistical properties of language structure rather than on absolute logical rules; and (ii) a focus on authentic data drawn from unrestricted domains of discourse rather than on invented examples. The two points are linked: the use of statistics is a consequence of the need for algorithms which are robust enough to cope with authentic rather than pre-selected linguistic material. Since the outside world has little use for CL [= computational linguistics; NO] unless it can handle authentic language, the future of the statistical approach seems assured."
>
> (1987a: 16)

Corpus linguistics has from time to time been subject to criticism regarding the use of corpus data. The criticisms that are heard find their origins in two different sources. One source is the misconception about the nature of today's use of corpus data, confusing the current role of corpus data with the role they were given in past approaches. A second source from which criticisms originate is the conviction held by linguists from different (theoretical) backgrounds that introspective data are the most reliable data.

With respect to the first type of criticism, it should be noted that, while the use of corpus data in linguistics is not at all new, the role of the corpus in current corpus linguistics differs considerably from that which it played in older approaches. For example, for the grammarians in The Great Tradition a corpus of texts constituted the main source of data. The grammars that were produced were text-based in the sense that the grammatical descriptions were exemplified by means of examples derived from a corpus of texts. This particular use of corpus data has rightly been criticized on account of the fact that one tended to restrict the scope of description to examples that were the result of mere chance-selections (mostly from literary sources).

The second point of criticism relates to the exclusive role that some of today's adversaries of the use of corpus data unfailingly associate with a corpus. While it is true that a structuralist like Harris stated quite categorically that the corpus was the only legitimate object of study in linguistic research, this view is not subscribed to by today's corpus linguists. Harris stated that

> "Investigation in descriptive linguistics consists of recording utterances in a single dialect and analyzing the recorded material. The stock of recorded utterances constitutes the corpus of data, and the analysis which is made of it is a compact description of the distribution of the elements within it."
>
> (Harris, 1951: 12)

Apart from declaring the corpus to be the sole source of data, Harris here advocated a purely inductive approach to descriptive linguistics. During the time of early transformational theory, however, when the balance of linguistics was swung by an interest in language competence and purely inductive methods were abandoned for deductive ones, the corpus data which had been central to structuralist practice and which also traditional grammarians had made extensive use of, were replaced as well. Textual data which could only provide an account of language

performance no longer held the interest of the linguists that adhered to the transformational approach. Instead, introspective data were considered to be the most (and quite often also the only) reliable data. Moreover, intuitions -- and the linguist's own introspections in particular -- were held to be the easiest data to obtain. Langendoen, for example, observed that

> "Native speaker intuitions make up the *entirety* of the data available to linguists; the use therefore of such intuitions is to replace ... the 'corpus' of grammarians such as Harris ... There is more than enough data which stares you in the face without having to go look for it with refined analytic tools."
>
> (1969: 405)

The linguist who uses himself as an informant in collecting data about the acceptability and interpretation of grammatical constructions, it was argued, knows what linguistic evidence he is looking for.[7] Linguistically naïve informants typically do not. In fact, the soliciting of data from informants is hindered by the effects that extraneous factors may have on the procedure.

The exclusive use in theoretical linguistics since the days of early transformational theory of intuitive data -- and introspective data in particular -- must be objected to as much as the exclusive role that the corpus played in, for example, the structuralist approach. For all intuitive data, that is introspective and informant data alike, the danger of self-fulfilling prophesies taking effect is not imaginary. The combined use of both the linguist's intuitions and a corpus of texts that characterizes current corpus linguistic practice provides a guard against the danger of having one's linguistic descriptive theory skewed by what Newmeyer (1983: 66) refers to as "too great a reliance on data collected in one particular way".

1.4 Computational tools for (corpus) linguistics

Given the different objectives that underlie the various computational linguistic approaches, it goes without saying that the requirements

[7] Cf. Newmeyer (1983: 61): "... when linguists ask themselves: 'Is such-an[d]-such an acceptable sentence of English?' they know *exactly* what they want."

made with respect to the computational tools by each of these approaches diverge. As the history of the computer in (traditional, theoretical and descriptive) linguistics goes back only a short way, the development of computer tools especially geared to linguistic interests has not yet advanced to a standard equal to the tools used for other applications. Shieber (1985: 189) remarks on this point that "in the natural-language-processing community, the usefulness of computer tools for testing linguistic analyses is often taken for granted. Linguists, on the other hand, have generally been unaware of or ambivalent about such devices." The significance of the role the computer can play in linguistic research, however, is beyond question. As Shieber (1985: 190-3) observes: "The computer constitutes a straitjacket in that it can force rigorous consistency ... [it] serves as a touchstone for verifying the correctness of a grammatical analysis ... [and it] serves as a mirror, objectively reflecting everything within its purview."

1.4.1 Design issues

(Corpus) linguistics requires the development of tools that will permit the automatic processing of data in such a fashion that the linguist need not be concerned with aspects of programming, computational efficiency, and the like. Preferably the tools should make it possible for the linguist to proceed with a minimum of interference, allowing him to concentrate on matters that are within his expertise. The way things stand today, however, in practice there will be a trade-off between what is desirable and what is practical. As yet there are no systems that combine, for example, a maximum of efficiency with optimum linguistic felicity. In designing a system for linguistic analysis a number of issues must be considered and weighed against each other. Among these are the following:

1. the system should be suited for linguistic analysis;

2. it should be easy to use; and

3. the system should be efficient.

In designing a system that is suited specifically for use in linguistics the main concern consists in creating an environment in which linguistic hypotheses may be tested and further developed. The most central

component of such a system is of course the parser: it not only holds the hypotheses formulated by the linguist, it also determines the course of the analysis process and the nature of the results. The question what parser to incorporate in the system must therefore be considered the most critical with respect to the functionality of the system as a whole.

Parsers may be of two kinds: they either take the form of computer programs (these we shall henceforth refer to as hard-coded parsers), or they are grammar-based. Parsers that fall within the first category tend to be very efficient since they allow us to make optimum use of the computational qualities of the programming language that is employed. Unfortunately, however, their procedural nature seriously hinders an adequate linguistic description in which static relations are described that hold between various objects or constituents. A further drawback of using a hard-coded parser resides in the fact that computer programs generally are notoriously difficult to interpret for anyone but the writer (and the machine). A grammar-based parser, on the other hand, is the result of the automatic conversion of a formal grammar by means of a parser generator.[8] It allows the linguist to express his hypotheses directly in terms of grammar rules. Consequently, he need not concern himself with aspects of computing. A further advantage of using a formal grammar rather than a computer program is that such a grammar can be considered interesting in its own right.

Over the years the use of formal grammars that has established itself in theoretical linguistics has found its way into descriptive linguistics. More and more research is devoted to developing grammar formalisms that not only have sufficient expressive power for the description of natural language (in general), but that also make it possible to write grammars for specific languages without being hindered by any restrictions imposed by the formalism regarding the use of a given linguistic descriptive framework. Although the choice of the grammar formalism is primarily dependent on what linguistic theory one subscribes to, it is further determined by the question whether or not a parser generator exists. Among the grammar formalisms that have been and continue to be used on quite a large scale are context-free grammar, augmented phrase structure grammar, augmented transition network grammar, and equivalents of these.

[8] A parser generator is a computer program that converts a formal grammar into a parser.

Irrespective of the grammar formalism that is opted for, any linguistic description is bound to show lacunae, whether they are phenomena or structures that were simply overlooked or things that one was unaware occurred at all. To compensate for shortcomings in the grammar, the system should be designed to operate interactively whenever this is desired. Moreover, apart from the flexibility that can be introduced into the system by having it operate interactively, its adaptability and functionality may be further enhanced by means of a modular structure.

Although generally speaking efficiency will be of the utmost importance only with application-oriented systems, while in testing a theoretical model it is hardly relevant at all, systems designed for corpus analysis cannot afford to be too inefficient. Since corpus linguistics deals with large amounts of data that have to be processed, it cannot afford the degree of inefficiency one can observe from time to time in theory-motivated systems. Yet, in practice, in the trade-off between efficiency on the one hand, and coverage, detail of analysis and correctness on the other hand, corpus linguistics tends to let the latter interest prevail. In opting for less than optimal parsers (by adhering to grammar-based parsers rather than hard-coded ones, while aiming at full coverage, detail of analysis and correctness), corpus linguistics puts efficiency in second place. Optimalizations in the parsing process are then sought in, for example, a reduction of the ambiguity by providing additional information through intervention.

Whereas linguistic felicity and expressiveness, as well as efficiency of the parser are relatively objective measures for evaluating a system, ease of use is not. Yet a system can be said to be easy to use when it does not provide much of a threshold to the linguist in computerizing his linguistic hypothesis, testing it, revising it, etc. Obviously, ease of use begins with transparency, user-friendliness and good documentation.

We observed earlier that the development of tools for doing linguistics presumes a weighing of priorities among the design issues. Depending on the objectives one holds some considerations will take priority over others. Consequently, systems may differ considerably. The systems that have been designed and implemented over the years fall into three main groups: they are application-oriented, theory-motivated, or corpus-oriented. Systems that fall within the first category tend to put efficiency first, are eminently robust and flexible. They are aimed in the

first place at obtaining results -- whether in interpreting questions and producing answers or analyzing the informational structure of a text. Although essentially linguistic components may be used, these systems are generally less linguistically oriented than theory-motivated or corpus-oriented systems. Theory-motivated systems, as the term already implies, are motivated by a particular linguistic theory. They make it possible to test (parts of) a given theory against a range of data. Moreover, such systems can be used in the development of theoretical models that are aimed at a psychologically realistic representation of the human language faculty. Finally, corpus-oriented systems are designed so as to allow for the processing of large amounts of language data. Since their principal users are linguists investigating language use, it is a matter of course that they should incorporate extensive and detailed linguistic components.

Below we include a discussion of three systems that were developed in the course of the years. All three have been applied in the analysis of English. However, each of them falls in a different category. Thus the natural language information processing system developed by the Linguistic String Project (LSP) is an example of an application-oriented system. The Parsifal system, on the other hand, is theory-motivated, while the TOSCA system is corpus-oriented.

1.4.2 The LSP system

The development of a natural language information processing system was undertaken by the Linguistic String Project at New York University as early as 1964-5. At that time a basic parser and grammar were designed and implemented that would form the basis of a full-fledged information processing system. This system was intended for the computerized analysis of (English) text and should make it possible to get access to and indeed retrieve information from any natural language material put to it. In 1981, some 15 odd years and several implementations later, Sager in a publication entitled *Natural Language Information Processing. A Computer Grammar of English and Its Applications* describes the system as it had then been developed. As it appears, the main focus of the work during this past period had been the processing of scientific articles and technical reports, especially in the field of medical science.

The system as Sager (1981) describes it consists of a number of

components. Apart from the basic grammar and parser these are[9]:

- a word dictionary which provides parts of speech and syntactic subclass information for each word;
- a programming language especially designed for writing natural language grammars;
- procedures for transforming syntactic parse trees into transformationally equivalent, less varied structures;
- programs for text and dictionary work (concordance, dictionary update, and other functions) and a clustering program that operates on grammatically analyzed sentences of a subfield to generate semantic word classes for the subfield.

The interaction between the various components in the processing of the material is as follows:

> "... the type of grammar used, linguistic string analysis, provides an analysis of each sentence into component word-strings that are at the same time both the grammatical and informational units of the sentence. Further operations refine and specialize the outputs of the string analysis program to obtain the desired informational representation of sentences. In this system, there is no call upon a special independent semantic component in order to achieve highly sophisticated information processing. The string parse tree and subsequent operations of transforming and labelling its components are the means of providing an informational characterization of the input texts."
>
> (Sager, 1981: 15)

The grammar consists of a set of context-free rules that are augmented with a set of conditions. These conditions, or restrictions as they are called, are stated in a restriction language. Restrictions may be of two kinds: they either serve to define certain well-formedness constraints that cannot conveniently be expressed in the context-free rules of the grammar, or they are used to optimize the parsing process. An example of the first type of restriction would be the test on subject-verb concord. Optimizations are generally achieved through look-ahead, making it

[9] Cf. Sager (1981: 14-15).

possible for the parser to pass over certain alternatives.

Apart from the restriction-type optimizations, the efficiency of the parsing process is further enhanced by the definition of nested subsets of the grammar. The smallest of these subsets describes the constructions that are supposed to be the more common in a language, while other, larger subsets also include less frequent ones. During the analysis process, a sentence is first analyzed on the basis of the rules contained in the smallest subset. Only when the analysis fails, are rules of the next larger subset brought into operation. An additional advantage of the use of such nested subsets is of course that the ambiguity that in an overall grammar arises from having to account for all constructions (whether frequent or infrequent), is reduced considerably.

The grammar (including the restriction component) that was written and implemented for scientific English is claimed to have a reasonably full coverage of English, while extending it should be easy. Yet the basic categories that are used in the grammar are not the basic categories commonly found in and suited for linguistic description, but they have been established in a purely deductive fashion on the basis of the material that is being investigated. Extension of the grammar, or writing a similar grammar for a different language altogether, may therefore not be as trivial as one is made to believe. Further drawbacks of this string analysis approach are that, from a linguistic point of view, the terminology used tends to be rather obscure and that analyses lack linguistically relevant generalizations.

1.4.3 The Parsifal system

The notion of deep structure underlying surface sentence structure as it is maintained within the framework of transformational grammar has also been used in the development of natural language processing systems. Using transformational grammars for analysis, however, proved problematic. The main stumbling-block was found to be the computation of inverse transformations, which appeared highly nondeterministic. In order to overcome this problem different strategies were followed. One consisted in obtaining an analysis through synthesis. On the basis of a given grammar all possible sentences would be generated and then matched against the input sentence. This enumerative approach was not only very expensive (computationally), it also was considered psychologically unrealistic. An alternative approach was found in a two-step procedure. This consisted in analyzing a given sen-

tence on the basis of a context-free grammar that accounts for a superset of the structures that are generated by the transformational grammar. Next, inverse transformations would be applied to the set of structures that had been obtained in the first step, looking for an analysis that would fit the base transformational grammar. Again in this approach the computational difficulties in computing the inverse transformations proved insuperable.

In the late 1970s Marcus[10] developed a system, Parsifal, that was based on the extended standard theory. The system was designed as a model of the human language faculty. It was intended to demonstrate that there is a theoretical significance to the determinism hypothesis, which consists in the assumption that human language processing is essentially equivalent to deterministic parsing. Marcus' parser assigns to each sentence that is entered a surface syntactic structure that is immediately related to the underlying deep structure that is postulated. The assignment of this structure, which includes both phrase boundaries and traces for deleted elements, no longer holds the problems that arose with the deep structure type of analysis that had been attempted earlier.

The object language is described by means of rule-packets consisting of situation-action type of rules. These are expressed in terms of "Pidgin", a restricted language that is formulated to look like English, and can be automatically converted into LISP. The parser uses two types of data-structure, an active node stack and a buffer. The active node stack is a stack of nodes for which daughters are being sought, while the buffer contains a sequence of nodes seeking attachment to their mothers. The buffer contains three constituents at most at any one time and thus provides limited look-ahead. Since the parser is deterministic it can be very efficient.

So far what is claimed to be a fairly complex grammar of English has been implemented. However, many syntactic phenomena have not yet been included, among which are coordination, pp-attachment and lexical ambiguity. Moreover, the system has not been applied to other languages. As it stands, the theory underlying the design of the system and the claims that go with it may yet be falsified. Meanwhile, large-scale versions of the parser are being developed for industrial application. Finally, the Parsifal system has been taken as the basis for a number of other systems (e.g. Paragram, see Charniak, 1981).

---

[10] The standard reference here is Marcus (1980).

1.4.4 The TOSCA system

The TOSCA system was developed at the beginning of the 1980s. It was especially designed for use in corpus linguistics. Until the mid-seventies corpus linguistics had been occupied primarily with the compilation of corpora, concordances, word-frequency lists and the like. As computers became available on a larger scale and the use of computer-readable corpora spread, the desire to go beyond mere word-based analyses and to include the level of syntax in the study of language use grew accordingly. However, the tools for doing this kind of 'advanced' corpus linguistics were lacking. This was the situation when the Dutch Computer Corpus Pilot Project (CCPP) was started in 1976. This project was initiated in Nijmegen and was participated in by most English Departments in Holland.

One of the objectives with which the CCPP was set up was to gain experience in using the computer in the syntactic analysis of corpora. In the course of the project a system was developed by means of which corpora that have been tagged for word-class categories and constituent-boundaries can be analyzed syntactically, using a context-free grammar. As the analysis of the 130,000 word Nijmegen Corpus would show, however, the system suffered from a number of shortcomings. The most important of these relate to the way in which linguistic knowledge is employed in the process of analysis. In the system the context-free grammar only plays an ancillary role; it provides the basically unlabelled bracketings as they have resulted from the preprocessing phase with function and category labels. While in the context-free grammar linguistic intuitions with respect to the higher level syntactic structure are formalized, intuitions about the interpretation of sentences and also about morphological rules and rules for word class assignment are not formalized at all. The tagging that was done in the preprocessing phase and in which this knowledge was employed, was carried out manually for lack of any other means. Despite the availability of a manual of instruction that the linguists involved in this process could consult, the procedure proved to be extremely error-prone and yielded numerous inconsistencies. Moreover, it was very time-consuming.

With the development of the TOSCA system, a system was envisaged that would make it possible to process large untagged corpora of texts in such a way that they will produce and save detailed information. The design of the system looks as represented in Figure 1.

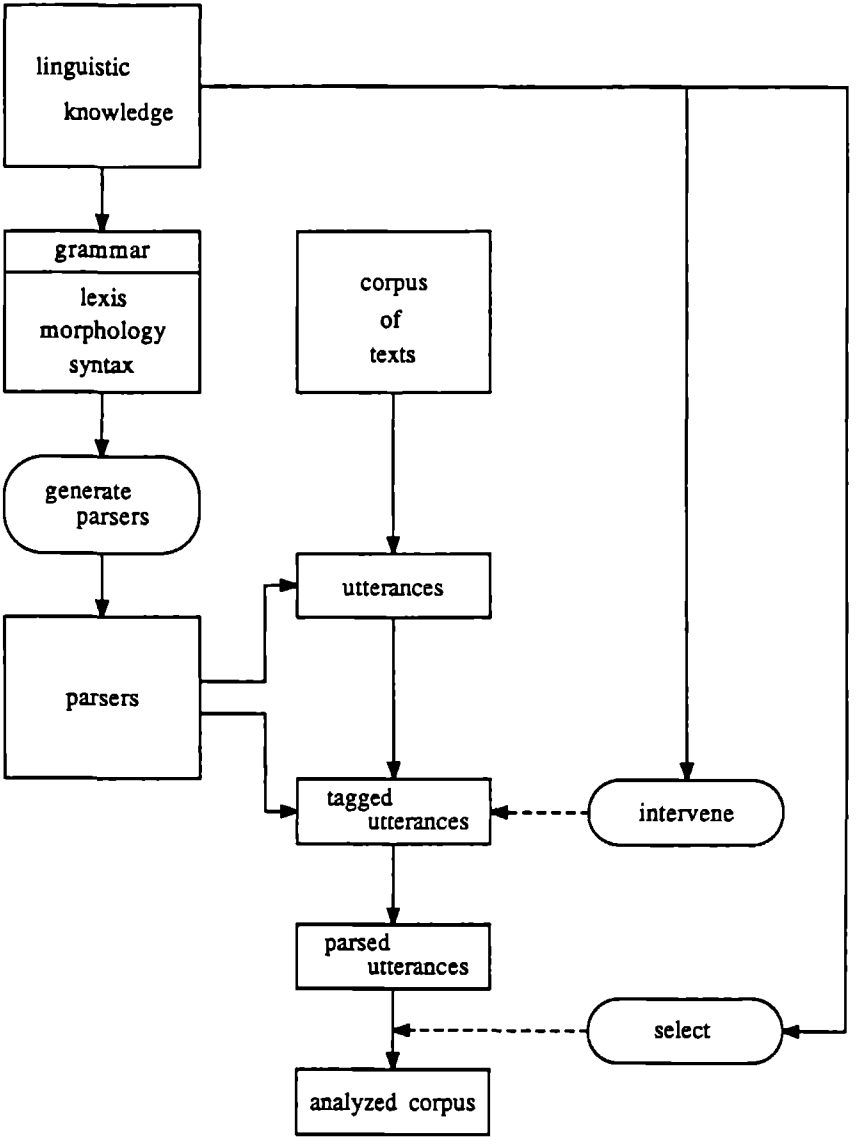


Figure 1: The design of the TOSCA system

The system itself is language independent in the sense that it consists basically of a grammar that operates on a corpus; however, it requires a set of language specific input components. The grammar is a formal grammar that lends itself to automatic parser generation, i.e. it can be automatically converted into a (number of) parser(s). The parsers are then used for the analysis of the utterances contained in the corpus. The analysis results, which have the form of trees representing the syntactic structure of the utterances in the corpus, are stored in a database.

Ideally the grammar should consist of a set of rules that account for all the structures encountered in a corpus. With such a grammar it would then be possible to have the syntactic analysis process run autonomously. Corpus linguistic practice, however, is a long way removed from this ideal situation. In writing a formal grammar the linguist relies very much on his intuitions about the form (and further possible variant forms) that various structures in the corpus language may assume. Although (prior to the actual analysis of the corpus) in a process of testing and revising the grammar, flaws in earlier hypotheses as laid down in the grammar may be corrected, we have found that a grammar is bound to be 'incomplete'.[11] Consequently, the analysis may fail and human intervention is required in order to supply the missing information. Another point where intervention is called for is in instances where the grammar yields numerous different analyses for a given string.

In order to overcome the problems that arise from the shortcomings of the grammar, an interactive system is a must. The TOSCA system was therefore designed in such a way that interventions can easily be made by the linguist. The role of interventions is twofold. They do not only serve to provide information that is lacking in the grammar, they also enable the analysis process to run more efficiently. A process in which interventions are allowed is liable to adopt ad hoc procedures. As a safeguard, we therefore restrict interventions to actions undertaken in reaction to some prompt given by the system itself.

In the TOSCA system interventions are made possible by making use of the modular structure of the grammar, or putting it more precisely, by making use of the modularity of the parsing sequence. The modules

[11] See also chapter 3 (section 3.3).

of the grammar are converted into a sequence of parsers. Interventions can be made at the interfaces between these parsers. Here the linguist can provide any missing information, or give direction to the analysis by selection.

When we consider the parsing sequence that we find in the TOSCA environment, we can distinguish the following parsing steps[12]:

tokenization

the tokenizer describes the text in a sequence of tokens, separates punctuation marks from the words to which they are appended, separates utterances, handles hyphenation, abbreviations, sentence-initial capitalization, etc. Tokenization is the first step in the parsing process; it precedes any dictionary look-up.

morphology

the morphological parser(s) tries/try to derive the tokens from a possible lexical form by means of regular morphological rules; it also suggests lexical categories for the words on the basis of their morphological features. Like the tokenizer, this component precedes and is independent of any dictionary look-up.

lexicon

all strings that have been suggested as possible lexical items by previous steps in the parsing process are now looked up in the dictionary and, where possible, provided with word class and feature information.

lemmatization

to each of the tokens a word class and feature set is assigned on the basis of the information that was acquired for these.

---

[12] Cf. van den Heuvel (1987: 239). Note that, with the exception of the first parser, which operates on the raw text material, all parsers take the data resulting from a preceding parser and transduce it into material with a similar format but with a richer structure.

syntax

> the syntactic parser operates on sequences of word classes and subsequently yields a labelled bracketing containing function and category information, for all constituents, ranging from immediate to ultimate constituents of sentences.

The output of the syntactic analysis is stored in the database where it is available for consultation.

The TOSCA system described here was developed in the course of the first of the two TOSCA projects. This project was aimed at the design and implementation of computational tools specifically geared to use in corpus linguistics (TOSCA = TOols for Syntactic Corpus Analysis).[13] In the course of the project not only the prototype of the system was developed, but there was also some experimentation with the type of formal grammar that could be used as input to the system. In close cooperation with members of the Computing Science Department of Nijmegen University, the formalism of Extended Affix Grammar -- which originated from the field of computer science where it was used for the description of artificial languages -- was applied and adapted to the description of natural languages.

The first TOSCA project was succeeded by a number of projects, each of which was concerned with the description and subsequent analysis of a specific language. The language-independent tools that had been developed thus came to be applied to Modern Standard Arabic (Ditters, 1987), European Spanish (Hallebeek, 1990), and Contemporary British English.

The project aimed at the analysis of a corpus of British English is the second TOSCA project. The primary objective in the analysis of the material under investigation is the study of linguistic variation. To this end a corpus has been compiled and a grammar written. These constitute the topics of the remaining chapters of this book.

---

[13] The Linguistic Database (van Halteren and Oostdijk, 1988; also van Halteren and van den Heuvel, 1990) in which the analyses are stored was developed in a separate project, which ran parallel to the first TOSCA project.

2 A Corpus Linguistic Approach to Linguistic Variation

2.1 Introductory

In the previous chapter corpus linguistics was characterized both as a branch of computational linguistics and as a formalized approach to descriptive linguistics. It was observed that corpus linguistics aims at the study of actual language use and that to this end the large-scale analysis of corpora is pursued. Since the previous chapter was concerned with contrasting corpus linguistics with other computational approaches, the focus of attention has so far almost exclusively been on the tools that are employed in the analysis process. Little attention has yet been given to the object of study: the corpus. Therefore, the present chapter is devoted to this topic.

Usually a corpus is understood to be a collection of texts which represent different kinds of language. As such it forms an ideal basis for the study of phenomena such as register, medium and style. In the process of studying language variation the analysis of a corpus is but a preparatory step. A detailed syntactic analysis gives access to the structures and their realizations contained in the utterances. This, in turn, provides a basis for the derivation of quantitative data, and the establishment of the nature of various linguistic variants. Next, preconceived notions about the relation between the variants and their linguistic and extra-linguistic determinants can be verified by comparing the characteristics of different language varieties as they are represented in the samples.

The conception of language as a complex of many different varieties is not at all new. Over the years linguistic subdisciplines have been concerned with (aspects of) linguistic variation. However, the complexity of the variability in language was thought to be unmanageable and to aim for an integral description of linguistic variation was considered to be unrealistic. Therefore, linguists in their descriptions have tended to abstract away from this variability. Studies of linguistic variation have been restricted to the setting up of small-scale hypothetical models which have contributed fairly few insights into the phenomenon of language variation.

In part the apparent failure to cope with the complexity of such a phenomenon as language variation can be attributed to the fact that

linguists have not been very well equipped to carry out large-scale formal empirical analyses which would enable them to systematically vary extra-linguistic factors and examine the accompanying linguistic variation. The present chapter[1] seeks to answer two main questions: 1) why, while obviously well-equipped for such an undertaking, corpus linguistics has so far failed to play any substantial role in the study of linguistic variation; and 2) what role it can play in the future, especially now that recent developments in the field of corpus linguistics have provided the means to get access to data that before could only be obtained on a very small scale or not at all, while at the same time, various techniques have been discovered and/or further developed which can be used in the manipulation and interpretation of data.

2.2 Language variation in linguistic theory

Over the years linguistic variation has appeared to be a problem area in linguistics. Aarts (1984), in discussing the attempts that have been made to come to terms with the problems that linguistic variation poses, reaches the conclusion that "for the time being it looks as if the description of language use requires idealization just as much as the description of language structure" (Aarts, 1984: 73). He also points out that "those who do not believe in idealization have made things very difficult for themselves. They have not yet proved that the construction of a complete and integrated 'variety grammar' of English is a feasible proposition" (Aarts, 1984: 72). Yet Aarts cannot deny that the integration of linguistic variation into linguistic theory is slowly making progress. Below we first consider the study of linguistic variation in the past. After that, attention is given to the potential role of corpus linguistics against the background of more recent developments: the advancement of computer technology on the one hand, and the development of corpus linguistics on the other. In this light we discuss the approach adopted by the second TOSCA project which has set out to investigate linguistic variation through a corpus of present-day English that was especially designed for this purpose.

[1] Parts of this chapter were originally published in *Literary and Linguistic Computing*, Vol. 3 No. 1, 1988: 12-25, and in *ICAME Journal*, Vol. 12, 1988: 3-14.

2.2.1 Background

It is a well-known fact that a language is not a homogeneous phenomenon but rather a complex of many different varieties. The existence of linguistic variation is something linguists have long been aware of. Yet in their descriptions they have tended to abstract away from this variability because, as Chambers and Trudgill point out, they "have started from the assumption that variability in language is unmanageable, or uninteresting or both" (Chambers and Trudgill, 1980: 145). If we look upon a language variety as "a sub-set of formal and/or substantial features which correlate with a particular type of socio-situational feature" (Catford, 1965: 84) we cannot but conclude that it is the infinite exhaustiveness of the situation that has led people to believe that attempts to incorporate linguistic variation in their descriptions are bound to fail.[2]

Linguistic theory has seen the introduction of notions like de Saussure's 'langue' and Chomsky's 'linguistic competence', both of which point to an attempt to start analysis at a more homogeneous level. A similar idealization can be observed in Modern English grammars, which are typically 'common core' grammars. These can be said to be variety-neutral since they are concerned with those linguistic features that the range of utterances in various varieties are assumed to share, regardless of any extra-linguistic dimensions. The extra-linguistic dimensions to language, the study of language use and its determinants, have been the concern of autonomous branches in linguistics such as stylistics and sociolinguistics, involving again an idealization of the data, in so far that these branches fail to integrate any extra-linguistic features that do not fall within their scope. If, however, we take it that linguistic theory should comprise a full description not only of linguistic competence but also of language use, we will have to go beyond such idealizations and try to come to terms not only with language structure but also with all its relevant extra-linguistic correlates.

---

[2] The introduction of variable rules in linguistic description has generally remained restricted to the phonological level, as for example in the work of Labov (1979, 1972). So far, the work that has been done with respect to the description of linguistic variation on the level of syntax has been rather fragmentary, so that, as yet, there are no linguistic descriptions incorporating variable, grammatical rules in which linguistic variants are related to their extra-linguistic correlates.

2.2.2 A descriptive framework for the classification of varieties

The awareness of the need of a theory of language use and, therefore, of linguistic variation has led to the development of descriptive models, or frameworks as Catford calls them, "of categories for the classification of 'sublanguages' or varieties within a total language" (Catford, 1965: 83). The models that have been developed so far demonstrate a complete lack of explicitness as far as the setting up of these categories is concerned. This is particularly true of the variety-based models which, simply assuming that a grammar can be written for any variety, fail to explicitly define the criteria on the basis of which varieties can be distinguished. The situation appears to be even worse, however, in the case of the item-based models. These, in which, as Hudson points out, "each linguistic item is associated with a social description which says who uses it and when" (Hudson, 1980: 40), are unmanageable because they are, by definition, open ended. In such an approach the setting up of categories is carried to the extreme, so that each item becomes a unique phenomenon. Since no basis is provided on which generalizations can be made, it is clear that the notion language variety is rendered inapplicable.

The conclusion to be drawn from the above is that as yet there are no satisfactory descriptive models for language variation. Therefore it is suggested that we reconsider the merits of the models that have been employed so far in order to see what aspects can be employed in a different approach.

Although the item-based type of model is best abandoned, there is one aspect that deserves our attention: such a model does allow for the matching of linguistic features with complexes of extra-linguistic determinants; linguistic items of a text need not (and usually do not) all correlate with one and the same extra-linguistic determinant.

Assuming that a grammar can be written for any variety, the variety-based type of model at least provides a basis on which we may proceed, in that it gives a (sometimes detailed) account of the range of extra-linguistic determinants that may be employed in the classification of varieties. In this respect the model proposed by Gregory offers some interesting categories.

2.2.3 Gregory's categories of varieties differentiation

The assumption underlying Gregory's model is that any text can be described in terms of features which correlate with the speaker by whom and the situation in which it was produced. In this respect it closely resembles contextualist models like the one proposed by Crystal and Davy (1969). In Gregory's opinion "a variety category can be thought of then as a kind of contextual category, correlating groupings of linguistic features with recurrent situational features" (Gregory, 1967: 178). He suggests that apart from the contextual categories of 'idiolect', 'temporal dialect', and 'mode', "it is also helpful to be explicit about and use in the description of language events aiming towards statements of meaning a separate, though related, set of situational categories for the description of those socio-situational features which may be expected to correlate with sub-sets of linguistic features" (Gregory, 1967: 178).

In setting up his variety categories Gregory distinguishes between those situational features in varieties distinction that relate to the reasonably permanent characteristics[3] of the user and those that relate to the recurrent characteristics of the user's use of the language. The former include the user's individuality, his temporal, geographical and social provenance, and his range of intelligibility. Their regular correlations with certain linguistic features leads to the establishment of the 'contextual categories of dialectal language variety', as Gregory calls them. The latter are categorized along three dimensions of diatypic variation. Thus we have the contextual categories of 'field', 'mode', and 'tenor of discourse' relating to the user's 'purposive role', 'medium relationship' and 'addressee relationship' respectively. Diagrams 1 and 2 display the suggested ca.egories of dialectal and diatypic language variety.

---

[3] Gregory prefers to speak of 'reasonably permanent' characteristics rather than 'permanent' ones "because although a user's individuality, temporal, geographical and social provenances, range of intelligibility within a community, all have a high degree of constancy, it is of course possible, as has already been suggested, for a language user to assume, at least partially, the linguistic habits of another individual, time, place and social class. Many English speakers control both a standard and non-standard dialect: the selection of one rather than another in different situations being closely linked with the question of use -- particularly of addressee relationship, the type of situation variation yielding linguistic DIATYPIC VARIETIES -- the linguistic reflections of the user's use of language in situations" (Gregory, 1967: 184).

DIAGRAM 1 (cf. Gregory, 1967: 181)
Categories of dialectal variety differentiation

	situational categories	*contextual categories*	*examples of English varieties (descriptive contextual categories)*	
user's	individuality	idiolect	Mr. X's English, Miss Y's English	DIALECTAL VARIETIES: the linguistic reflection of reasonably permanent characteristics of the USER in language situations
	temporal provenance	temporal dialect	Old English, Modern English	
	geographical provenance	geographical dialect	British English, American English	
	social provenance	social dialect	Upper Class English, Middle Class English	
	range of intelligibility	standard/non standard dialect	Standard English, Non Standard English	

DIAGRAM 2 (cf. Gregory, 1967: 188)
Categories of dialectal variety differentiation

	situational categories	*contextual categories*	*examples of English varieties (descriptive contextual categories)*	
user's	purposive role	field of discourse	Technical English, Non-Technical English	DIATYPIC VARIETIES: the linguistic reflection of recurrent characteristics of user's USE of language in situations
	medium relationship	mode of discourse	Spoken English, Written English	
	addressee relationship	tenor of discourse		
	(a) personal	personal tenor	Formal English, Informal English	
	(b) functional	functional tenor	Didactic English, Non-Didactic English	

So far Gregory's subcategorization is fairly straightforward, although the category of 'standard dialect' is presented, as Gregory himself points out, rather tentatively. 'Standard' here does not refer to any particular geographical or social provenance, rather it serves to indicate what Abercrombie (1955: 11)[4] has called 'the universal form of language': "what enables, for example, certain users of English throughout the English speaking world to communicate intelligibly with each other."

Having set up the main categories, of which those of diatypic variation, he grants, must be regarded as highly hypothetical, Gregory proceeds by exemplifying in what way one may arrive at a useful subcategorization for each of the contextual categories of 'field', 'tenor', and 'mode'. Of these only the last is sub-categorized in such a way that it opens up perspectives of coming to grips with this category and the complex of variables by which it is determined. This subclassification can be found in Diagram 3.

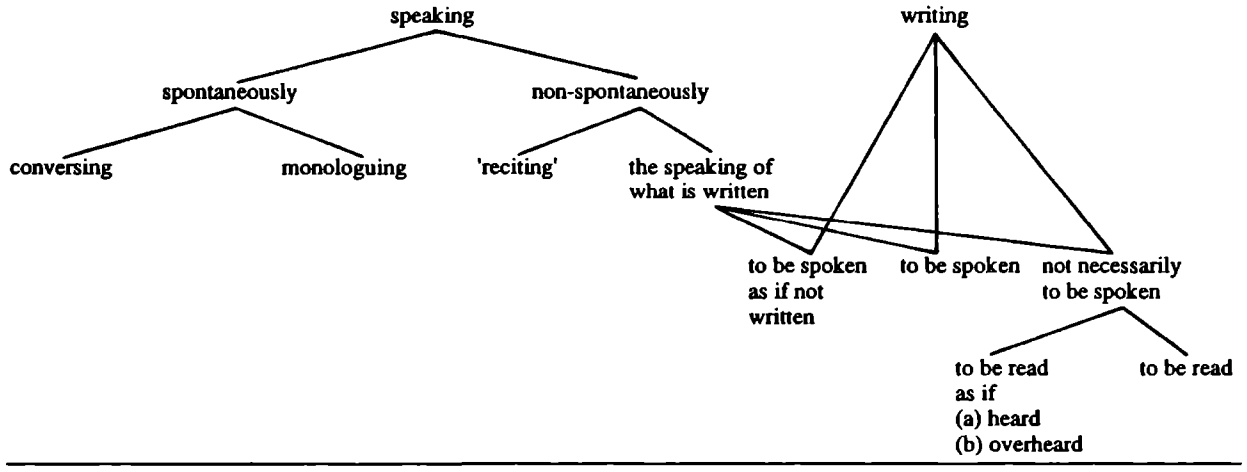
Whereas Gregory provides a quite detailed diversification in the subcategorization of the contextual category of 'mode of discourse', the suggested distinctions along the dimensions of situation variation categorized as 'the user's purposive role' and 'the user's addressee relationship' remain vague and lack precision.

It is here that Gregory suggests (particularly in the case of 'tenor of discourse') that at a more advanced stage in varieties description "more descriptive realization of the other individual dimensions might lead to the discernment of at least two (as here) and possibly several related subdimensions" (Gregory, 1967: 188). Language use may thus be found to vary with respect to the degree of formality (from extremely formal to particularly informal), depending on the relationship between addresser and addressee(s), whereas it may also be observed to correlate with the functional relationship that exists between them thus leading to the discernment and establishment of such sub-categories as 'didactic' vs. 'non-didactic', 'expository' vs. 'non-expository', etc.

Gregory's failure to substantiate his sub-categorizations and his appeal to further "more sophisticated" investigation confirm once more what was apparent in earlier findings and what led Crystal and Davy to conclude (in 1969) that they had reached a stage where they "would do well to wait for practical analysis to catch up, so that the theoretical cat-

[4] See also Gregory (1967: 183).

DIAGRAM 3 (cf. Gregory, 1967: 189)
Distinctions along the dimensions of situation variation
categorized as user's medium relationship



egories may be tested against a wide range of data, and more detailed analysis of texts carried out" (Crystal and Davy, 1969: 62).

2.2.4 The present state of affairs

The observation made by Crystal and Davy way back in 1969 still holds. Even in 1990, no one had "described the full range of linguistic correlates of any one of the dimensions; nor has there been much experimentation -- such as systematically varying the extra-linguistic factors and examining the accompanying linguistic variation" (Crystal and Davy, 1969: 65). Models have been developed, modified and/or rejected, but eventually all that has in fact remained is an awareness that the study of language variety will never proceed beyond this hypothetical stage unless some sort of large-scale formal empirical analysis is embarked upon. It is therefore all the more surprising to find that the possibilities of employing the computer in linguistic analysis, as was suggested by for instance Ellis and Ure (1969), have so far only been explored rather reservedly.

2.3 A corpus linguistic approach

In the preceding section it was pointed out that linguistic variation can only be studied through vast amounts of data. Moreover, it was observed that some sort of large-scale formal empirical analysis should be embarked upon if ever the study of linguistic variation is to proceed beyond the hypothetical stage. A branch of linguistics which by definition works with large amounts of data is corpus linguistics. Here large bodies of text are used as primary data in the study of actual language use. The use of the computer in the processing of the data is a necessary prerequisite, requiring, at the same time, the formalization of various techniques. It would seem therefore that as a branch of linguistics corpus linguistics is best equipped for the study of linguistic variation. Looking at past research, however, we cannot but observe that so far it has failed to play any substantial role in this respect. The reasons for this failure are the following:

1. the corpora that were used were not suited for studying linguistic variation;

2. computer methods had not yet reached the stage where it would allow for non-trivial information to be retrieved.

Some of the corpora used in corpus linguistics dated from the period in which material was collected on slips of paper and kept in filing cabinets, for anyone there to consult, but, unfortunately, not accessible through the computer. (An example is the *Survey of English Usage* which is discussed below.[5]) Other corpora had been compiled with the intention of representing a cross-section of either American or British English (the Brown Corpus and the LOB Corpus respectively). Samples of 2,000 words each were randomly selected from a wide range of texts. In selecting a great many different samples it was attempted to neutralize any variety-specificity. Exceptions to these were what can be referred to as 'specialized corpora'[6] which tend to be restricted to a (very) small subset of the language. Such specialized corpora could well be used in a variety study, given the availability of corpora exemplifying other subsets.

So far corpus-based variety studies could only make use of quantitative data available from a word-based analysis of the corpus, such as the frequency of occurrence of words or the mean sentence-length. The study of linguistic variation can, however, only seriously be undertaken if also ample quantitative data are available about the frequency of occurrence of syntactic structures. A further handicap experienced in earlier studies was that the statistical techniques were by no means as sophisticated as required for dealing with such complex data. It is only through recent developments in the field of corpus linguistics, such as the implementation of systems for the automatic analysis of text corpora and the development of sophisticated quantitative techniques, that the study of linguistic variation can now be considered a feasible

---

[5] Looking back Kaye (1987: 3) observes that in "1959 computers were slow and their storage and availability extremely limited; in fact many British universities did not have one. The decision to use hard copy was obviously correct, and remained so until this decade." Therefore, in the mid-eighties a project was started which aimed at the conversion of the written Survey data and their associated grammatical codings into a computer-based system.

[6] An example of such a 'specialized' corpus is the Louvain Drama Corpus which consists of samples of 62 British English plays written between 1966 and 1972. The corpus comprises 1,312,860 words.

proposition.

In the light of what was observed above it may be useful to have a look at the various corpora that have been used in English language research so far and consider them on their merits as far as the study of linguistic variation is concerned.[7]

2.3.1 Corpora in English language research

The Survey of (Educated) English Usage (SEU)

In 1959 Quirk initiated the large-scale undertaking of collecting material which he hoped would facilitate the identification of what he referred to as the "central core of educated usage".[8] Quirk observed the lack of a proper basis for the writing of a grammar of English and claimed that such a basis should be formed by samples of texts taken from the full range of varieties and strata of educated English, spoken as well as written. Each sample would have to be described in terms of its grammatical features, distinguishing between regular and variant forms of particular constructions, and the co-occurrence restrictions under which they appear. In addition to this linguistic description each text must be marked for its extra-linguistic variables, indicating the extra-linguistic occurrence restrictions for a given construction. Thus, Quirk claims, "the data assembled from the examination of this 'primary material', organized as a Descriptive Register, will provide adequate information

---

[7] We restrict ourselves to corpora that are generally available thus excluding corpora like the ones compiled in the Birmingham COBUILD project for purposes of lexicography. The COBUILD corpora comprise a vast amount of text, some 20 million words in all, and could well have been used for the study of linguistic variation. Unfortunately, however, distribution of the corpora appears impossible since the texts are under copyright. Moreover, under the British Information Act distribution to other countries in the world is prohibited.

[8] Quirk (1968: 79); Quirk adopts what he calls a 'working definition' which defines 'educated English' as English that is recognized as such by educated native speakers. He points out that, even though it may seem rather circular, this definition may be supported by results yielded by reaction tests. Furthermore, he claims that the use of the term 'educated English' as opposed to 'standard English' clearly draws attention to the (rather instable) social basis on which concepts like the latter rest.

for precise, objective, and comprehensive statements to be made describing the *majority* of English constructions, and the conditions under which they and their variants occur naturally" (Quirk, 1968: 79).

The compilation of this 'descriptive register' was guided by three main principles:

1. The primary material consists of *all* the grammatical data in samples of actually recorded English, spoken as well as written. Statistical data on the frequency of occurrence and its distribution of a construction and its variants will provide an objective basis for setting a norm and explaining departures from that norm.

2. Samples were to be taken from all varieties of educated English so as to give a representative picture of the full range of educated English usage.

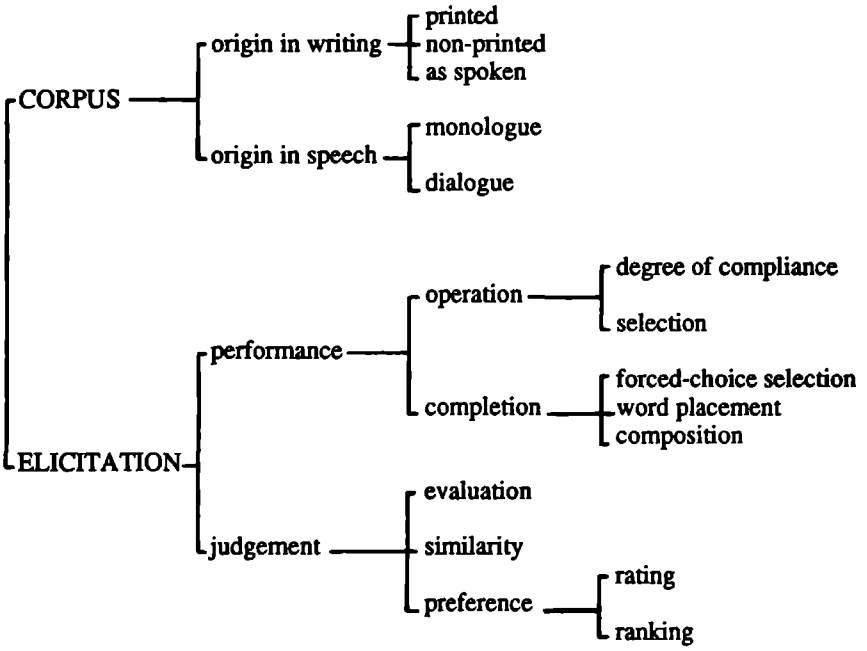
3. The Survey is concerned only with present-day (1950-) British English.[9]

The (primary) material that was collected consists of samples of 5,000 words each, taken from unscripted speech, novels, plays, poetry, criticism and other non-fictional prose, psychology and social sciences, law, politics, religion, 'useful arts' (e.g. cookery), newspapers, and so on.

Apart from this primary material some interpenetrating sampling was to be taken into account so as to determine whether for a given construction the saturation point had yet been reached. For high-frequency constructions it will not be difficult to collect sufficient material; for low-frequency constructions, however, and also for some variants of high-frequency ones, it will be necessary to provide supplementary samples. Therefore elicitation tests with native speakers were seen as an essential tool for enlarging upon corpus-derived information and for investigating features that might not occur in the corpus at all (Greenbaum, 1984: 193-201). The Survey as a whole may be summa-

[9] Quirk points out that taking 'present-day' to mean 'since 1950' is "a working rule made in the full realization that no arbitrary time-limitation will ensure absolute homogeneity and that even on the same day an educated man of sixty-five and an educated woman of twenty-five may differ in their usage; it is hoped that the chief linguistic variations occasioned by such factors as these, too, will be revealed at the stage of explanatory analysis" (1968: 80).

Table 1: Survey Strategy



(The design of the corpus can be found in Appendix A.)

rized as in Table 1, taken from Quirk and Svartvik (1979: 208).

Quirk claims that the Survey can play an important role in the description of what is actual and normal in linguistic behaviour, allowing for rules to be formulated on the basis of "the patterns that may be seen emerging from a corpus of natural material, in which at the same time the co-occurrent factors may be observed and from which statements may be made not merely listing but *ranking* the factors conditioning variants of these patterns" (Quirk, 1968: 81).

Although an approach as pursued in the Survey will be rather successful in making an inventory of the constructions used in present-day English, it is doubtful whether it will be possible to obtain any insights in the correlations between the linguistic features and the various extra-linguistic variables which have been used to describe the conditions under which they occur. A formal description of the corpus is

lacking. In our experience this may lead to inconsistencies in the analysis of the material.

The London-Lund Corpus

The Survey of Spoken English, a sister project of the Survey of English Usage, was carried out at Lund University. It was started in 1975. Its primary aim was to make available, in machine-readable form, the spoken material which had been collected in the Survey of English Usage. A secondary aim was to make the corpus available for linguistic research.

The computerized version of the spoken material found in the SEU is commonly referred to as the London-Lund Corpus. The Corpus includes all texts with their origin in speech. (For a listing of the contents of the material the reader is referred to Appendix B.) The material in this corpus differs from that found in the filing cabinets of University College London (being the original texts on slips of paper together with their transcriptions and prosodic analyses) to the extent that the number of features involved in the prosodic analysis has been reduced. Basic prosodic distinctions of tone units, nuclei, boosters and stresses have been retained, whereas other features, such as tempo, loudness, modifications in voice quality, and voice qualifications, have been omitted. Computer processing of the corpus has produced texts, concordances and word-lists. The corpus is being used for various studies on, for example, grammatical tagging, reference, questions and responses, negation, turn-taking, and interruption (Svartvik et al., 1982; Svartvik (ed.), 1990).

The Brown Corpus

The Standard[10] Corpus of Present-Day Edited American English is a one-million word computer-processible corpus of language texts assembled at Brown University during the years 1963-1964, and it is

[10] Kucera and Francis (1967: xvii) point out that the term 'standard' "is not intended as a qualitative description of the texts included. Rather, it is an expression of the hope that the Corpus ... may serve as standard of comparison for a variety of studies and analyses of present-day English."

therefore usually referred to as 'the Brown Corpus'. In their introduction to *Computational Analysis of Present-Day American English* (1967: xvii-xxv) Kucera and Francis describe the aim of the compilation of the Corpus as follows:

> "... the aim has been to compile a corpus of printed American English and present a basic analysis of the data according to the following criteria:
>
> 1. Definite and specific delimitation of the language texts included, so that scholars in using the Corpus may have a precise notion of the composition of the material
>
> 2. Complete synchronicity; texts published in a single calendar year only are included
>
> 3. A predetermined ratio of the various genres represented and a selection of individual samples through a random sampling procedure
>
> 4. Accessibility of the Corpus to automatic retrieval of all information contained in it which can be formally identified
>
> 5. An accurate and complete description of the basic statistical properties of the Corpus and of several subsets of the Corpus with the possibility of expanding such analysis to other sections or properties of the Corpus as may be required."

In the selection of the samples that make up the Corpus it was ascertained that the Corpus was fully synchronic, representative of a wide range of styles, and accurate. So as to ensure synchronicity the samples were all taken from material first printed in 1961. Further restrictions included those made with respect to the origin of the text (it had to be printed in the USA) and the author (so far as could be determined he had to be American), whereas the amount of dialogue in a selection had to be less than 50 per cent. Representativity was assured by random sampling of 500 samples of 2,000 words each, distributed over 15 categories, according to Kucera and Francis, representing "the full range of subject-matter and prose styles". The text categories can be found below in Table 2, in which also their distribution in the British English counterpart, the LOB Corpus, is displayed. For a listing of the contents of each of the major text categories the reader is referred to Appendix C.

The basic analyses that have been made of the Brown Corpus include (apart of course from the absolute and relative word-frequency counts) analyses of the distribution of occurrence of frequent words, the word-frequency distribution, word-length and sentence-length distribution in the Corpus. These are briefly discussed below.

The distribution of occurrence of frequent words

The information yielded by the range figure was felt to be insufficient in the case of high-frequency words since they occur in all genres and nearly all samples.[11] It was therefore decided to investigate the relative frequency of occurrence of each word in the individual genre subdivisions. The outcome of this study led Kucera and Francis to conclude that the uneven distribution among the subdivisions could not be attributed to chance but rather should be explained in terms of style and content characteristics of the genres significantly affecting the frequency of occurrence of even the most frequent words in English.

Word-frequency distribution

The frequency distribution shows some interesting facts at the 2,000 word level. Thus the proportion of hapax legomena (compared to the total vocabulary represented) appears to vary depending on the size of the sample. It is also suggested that the relation between type-token ratio and sample size might well be attributed to what is referred to as 'qualitative genre influences'.

Word-length distribution

An analysis was made of the length of all the words in the Corpus, the length of a word being defined as the number of graphic characters composing it. Basic information was collected about the graphic com-

[11] In Kucera and Francis (1967: 275) range figures specify in how many genres and in how many samples of the Corpus each word-type actually occurs. According to Kucera and Francis these "range indications are of importance in evaluating the significance of the frequency of occurrence, particularly in the case of middle- and low-frequency words."

position of both the dictionary of the Corpus as well as the running words in the stripped version of the Corpus, i.e., the version of the Corpus without any coding.

Sentence-length distribution

The study of sentence-length distribution in the Brown Corpus was undertaken by Marckworth and Bell. In Kucera and Francis (1967: 368-405) they report on an effort

1. to establish the sentence-length distribution of the whole population of the English Corpus and of each of the genres of which it is composed; and

2. to determine whether sentence-length is a significant parameter in the quantitative description of writing style in the various literary prose genres of the Corpus; i.e., does the genre impose some sort of constraint in this matter on the individual practitioner?

A comparison of the genre distributions shows that sentence-length is a measurably significant variable of genre style, its distribution being highly dependent on the classification of the genre as informative or imaginative prose. Marckworth and Bell rather speculatively suggest that it may well be possible that sentence-length distribution "is subtly dependent upon the expected relationship between author and audience, the nature and/or purpose of the information being conveyed, and the expected patterns set by previous examples of the genre" (Marckworth and Bell, 1967: 375).

Another outcome of this study is that the variety of different sentence-lengths occurring in a genre proved to be a significant parameter of genre. Marckworth and Bell conclude:

> "This factor may provide an index to the number of other stylistic variables, chiefly grammatical, that are used by an author within the confines of a genre. Genres can be ranked by degree of internal homogeny of sentence-length distribution patterns. This parameter is particularly useful in distinguishing differences in stylistic patterns between members of the imaginative prose category."
>
> (Marckworth and Bell, 1967: 375).

The LOB Corpus

The Lancaster-Oslo/Bergen (LOB) Corpus was assembled so as to provide a British English counterpart to the American Brown Corpus. In order to yield a corpus comparable to the Brown Corpus, which would facilitate a combined use of the two corpora, sampling methods and principles were closely followed. This resulted in a corpus of 500 British English text samples of about 2,000 words each. As in the Brown Corpus, the year of publication was 1961 and attempts were made to limit the amount of dialogue to 50 per cent or less. In the case of the LOB Corpus the author had to be British, non-British authors were excluded.

In Table 2 the basic composition of both the British and the American corpus is displayed (cf. Hofland and Johansson, 1982: 2; also Johansson, 1978: 3). For a listing of the contents of each of the major text categories in the LOB Corpus the reader is referred to Appendix D.

Text categories	number of texts in each category	
	Brown	LOB
A Press: reportage	44	44
B Press: editorial	27	27
C Press: reviews	17	17
D Religion	17	17
E Skills, trades, and hobbies	36	38
F Popular lore	48	44
G Belles lettres, biography, essays	75	77
H Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30
J Learned and scientific writings	80	80
K General fiction	29	29
L Mystery and detective fiction	24	24
M Science fiction	6	6
N Adventure and western fiction	29	29
P Romance and love story	29	29
R Humour	9	9
	500	500

Table 2: The basic composition of the LOB and the Brown corpora

The analyses that were made on the basis of the LOB Corpus, i.e. the ranking of the complete corpus and the comparison of the word frequencies occurring in different types of texts, do appear (as was expected) to correlate with the results yielded by the Brown Corpus. Hofland and Johansson confirm the fact that there are considerable differences between text categories even with the most frequent words. Like Kucera and Francis, this leads them to distinguish between two major groups of texts: informative and imaginative prose (text categories A-J and K-R respectively).[12] In addition to this major division Hofland and Johansson suggest that the contrast between fiction and non-fiction may be bridged by categories that could be termed 'essayistic prose' (text categories F, G, M, and R). On the basis of this new clustering they arrive at the following grouping of the text categories:

A-C (88 texts) : newspaper text
D-H (206 texts) : miscellaneous informative prose
J (80 texts) : learned and scientific English
K-R (126 texts) : fiction

Tagged versions of the Brown Corpus and the LOB Corpus

Apart from the original untagged version of the Brown Corpus a tagged version is available in which each word in the Corpus has been supplied with a code indicating its place in a taxonomy based on surface syntactic function. The tags are of five kinds:

1. major form-classes ('parts of speech'): noun, common and proper; verb; adjective; in short, the open lexical classes;

2. function words: determiners, prepositions, conjunctions, pronouns, etc.; the closed lexical and grammatical classes;

3. certain important individual words: *not*, existential *there*, infinitival *to*, the forms of the verbs *do*, *be* and *have*, whether auxiliaries or full verbs;

4. punctuation marks of syntactic significance;

5. inflectional morphemes, notably noun plural and possessive; verb past, present and past participle, and 3rd person singular

[12] Informative prose is also referred to as 'non-fiction', imaginative prose as 'fiction'.

> concord marker; comparative and superlative adjective and adverb suffixes.
>
> (Francis, 1980: 200)

A tagging of the material contained in the Brown Corpus was found to be desirable for a number of reasons. On the basis of the untagged material it was impossible to produce frequency tables in which homographs were disambiguated and inflectional variants brought together through lemmatization. The frequency tables would be much more informative if the entries were disambiguated and lemmatized. Moreover, a tagging of the material was seen as a necessary step in making the Corpus amenable to syntactic analysis.

For the same reason that had motivated the tagging of the Brown Corpus the grammatical tagging of the LOB Corpus was undertaken. A further reason was found in the fact that such a tagging would make it possible to compare the tagged versions of the two corpora. Therefore, in order to ensure the general comparability with the tagged Brown Corpus, largely the same set of tags was used that had been used for the Brown Corpus. (For practical reasons a number of tags were added to the Brown tag set, see Leech et al., 1983.)

2.3.2 Variety studies on the basis of the main corpora

It has already been observed that corpora which have been compiled with the intention of representing a cross-section of the language are not suited for the study of linguistic variation since, in selecting a great many different samples, they neutralize any variety-specificity. It is surprising to find (as we saw above) that attempts have been made to study linguistic variation on the basis of such corpora.[13] It is even more surprising when one realizes that until recently there was no access to anything but word-based analyses. Two main points of criticism can be made here:

1. Studies in linguistic variation on the basis of the main corpora (Brown, LOB and London-Lund) have failed to recognize that 'genre' is not a well-defined concept. The genres that have been

[13] Studies of linguistic variation on the basis of the Brown Corpus and the LOB Corpus include Ellegard (1978) and Johansson (1979).

distinguished so far have been identified on a purely intuitive basis. No empirical evidence has been provided for any of the genre distinctions that have been made;

2. in these studies it has, without any reservation, been assumed that correlations that were found to be significant on the word-level can be generalized to a full spectrum of linguistic features.

One of the most recent studies in linguistic variation is a study by Biber and Finegan (1986). Although this study is subject to the same criticisms that were made above with respect to earlier studies involving the main corpora,[14] it forms an important contribution in that it not only introduces the concept of 'text typology' but also develops a methodology for dealing with linguistic variation. Therefore a discussion is now included of some aspects of this study.

An initial typology of English text types

In their paper Biber and Finegan report on the Multi-Feature/Multi-Dimensional (MF/MD) approach to linguistic variation they have developed and which, they claim, "is particularly well-suited to the development of a typology of texts" (Biber and Finegan, 1986: 20). In defining the objectives of such a typology they recognize the need to a) identify the set of major text types in English and b) specify the relations among and between genres and text types. According to their working definition, genre categories are used to characterize texts on the basis of external criteria, while text types are defined in terms of linguistic characteristics of the texts themselves. It is clear that text types may represent groupings of texts that are similar with respect to their linguistic form and that do not belong to one and the same genre.

Biber and Finegan's idea of a text typology comes very close to Gregory's set of variety categories. Where Gregory described a variety category as "a kind of contextual category, correlating groupings of linguistic features with recurrent situational features" (Gregory, 1967: 178), Biber and Finegan point out that in developing their typology

[14] Biber and Finegan fail to evaluate the applicability of the genre distinctions that were made in the corpora they use as basis for their investigations. They simply adopt the genres as they were identified for the corpora. See also below.

they interpret the co-occurrence patterns among linguistic features and the relations among text types in functional terms. "Thus", they claim, "the resulting typology provides not only a mechanical classification of texts, but also a rubric for situating texts along each of several functional dimensions, with respect to salient contextual, cognitive, and social parameters of the text 'situation'." (Biber and Finegan, 1986: 20) The MF/MD approach Biber and Finegan developed and which they use in their research is characterized by:

1. the use of computer-based text corpora, providing a standardized data base and ready access to a wide range of variation in communicative situations and purposes;

2. the use of computer programs to count the frequency of certain linguistic features in a wide range of texts, enabling analysis of the distribution of many linguistic features across many texts and text types;

3. the use of multivariate statistical techniques, especially factor analysis, to determine co-occurrence relations among the linguistic features, facilitating the identification of underlying textual dimensions.

(Biber, 1985: 340)

The use of computer-based 'standardized' corpora[15] they advocate is essential since these exemplify much of the variation that must be accounted for. Moreover, such use enables replication of previous studies and allows for a comparison of the results with other studies of this kind. In their approach a factor analysis groups together linguistic features that co-occur with high frequency in each text. Next the factors are interpreted as textual dimensions, through the association of (groups of) texts with various contextual (situational, social, etc.) variables. Further steps in this approach include the computation of factor scores for each factor and each text, and the subsequent analysis of the distribution of the factor scores among the genres and, in the light of this, a further interpretation of the textual dimensions. Finally, texts are

[15] It is not entirely clear what Biber and Finegan mean by the term 'standardized corpora'. They use it to refer to corpora in which texts have been selected from "many of the possible genres of English" (Biber and Finegan, 1986: 25). It would seem therefore that it is meant to refer to those corpora that are intended as representations of cross-sections of the language.

clustered that are most similar to each other with respect to the dimensions distinguished. These clusters are then interpreted as "underlying text 'types', through assessment of the communicative parameters (situation, purpose, etc.) most widely shared by the texts grouped in each cluster" (Biber and Finegan, 1986: 24).

In their research Biber and Finegan have so far used the Brown Corpus, the LOB Corpus and the London-Lund Corpus. In the paper discussed here they report on a study they carried out on the basis of the LOB Corpus and the London-Lund Corpus as well as a small collection of professional letters, comprising one million odd words in all. The 545 texts they include are divided over 16 genres. The linguistic features they use have been identified in previous research as 'functional markers of different styles, modes or registers'. A factor analysis leads them to distinguish three textual dimensions along which texts are found to vary linguistically: Interactive vs. Edited Text, Abstract vs. Situated Content, and Reported vs. Immediate Style. Subsequently, with the help of a cluster analysis texts are grouped together that are maximally similar to each other so that the clusters can be interpreted as text types, i.e. groupings of texts that are similar in their linguistic form, regardless of external criteria. This results in an initial text typology in which nine text types are distinguished; these are

1. immediate interaction
2. formal exposition
3. informational-interactional text
4. present reportage
5. informal informational narrative
6. informal exposition
7. interactional narrative
8. informal exposition with narrative
9. imaginative narrative

A breakdown of the texts in each cluster by genre shows that texts from a single genre occur in different text types. Only in a number of cases can a majority of texts from a single genre be found in one particular cluster.

In their conclusion Biber and Finegan note that the typology they propose begins to show the complexity of the notion 'text type'. Their study, therefore, is preliminary: "a complete typology of text types will require inclusion of a fuller range of linguistic features and texts,

micro-analyses of individual texts, and a fuller discussion of the interaction among genres and text types" (Biber and Finegan, 1986: 41).

As Biber points out in an earlier paper (1985), an adequate data base and an adequate sampling of linguistic features are necessary prerequisites to the type of analysis they propose. With respect to the selection of texts he observes that

> "In selecting the texts to be used in a multi-feature/multi-dimensional approach, care must be taken to include a broad range of the possible situational, social, and communicative task variation occurring within the domain to be analyzed. This entails considerable preliminary research to identify (1) the parameters of situational variation within this domain; (2) the different processing constraints within this domain; (3) the different communicative tasks in this domain; and (4) the different relationships among communicative participants."
>
> (Biber, 1985: 341f)

The selection of linguistic features must, owing to the lack of other ways to identify potentially important linguistic features, be based on previous research. Here Biber is careful to note that at this stage, until sufficient analyses of this type have been carried out, research must be considered exploratory.

Whereas we agree with Biber and Finegan that the MF/MD approach may prove valuable in developing a typology of text types, some critical observations should be made.

Like others before them Biber and Finegan in selecting their texts, make use of material contained in corpora that have been compiled with the intention of representing a cross-section of the language. This means among other things that it is quite possible that their samples may yet be proven to have been too small. Moreover, one may question the extent to which certain samples can be regarded as representative of a particular genre, especially with those samples which come within a genre where numerous texts contain a large amount of dialogue. Here although generally random sampling procedures were used in the compilation of the corpora, samples were discarded if the amount of dialogue they contained exceeded 50 per cent. Another point of criticism concerns the fact that in their study Biber and Finegan never once question the validity of the genre distinctions they use. At first sight it might appear that this is hardly of any consequence, yet in

obtaining any insights in the relation between linguistic variation and the varying of extra-linguistic conditions we need to know what extra-linguistic variables we are dealing with. It may well be that the genres they distinguish are rather heterogeneous groups of texts which will be of little use in explaining the relations between linguistic and extra-linguistic variance. Defining a genre as an a priori text classification based on situation and purpose they ignore other possibly important factors, such as for instance the influence of personal style, topic and text-length.

In selecting the linguistic features for their study Biber and Finegan restrict themselves to linguistic features that have been identified in previous research as potentially important. The features considered so far include both lexical and syntactic features. The syntactic features, however, only comprise syntactic categories like if-clauses, wh-clauses, it-clefts, place adverbs, and time adverbs. It should also be noted that they find it impossible to automatically recognize all possible variants of a particular construction, whereas others cannot be recognized at all.[16] The question that remains to be answered is to what extent the features that are distinguished may be expected to cover the full range of possible linguistic variation.

2.3.3 Perspectives for corpus-based variety studies

As was observed earlier, corpus-based studies of linguistic variation have so far failed to make a substantial contribution to the development of a descriptive theory of linguistic variation. This failure, as we have seen, must be attributed in part to the lack of proper data and in part to

---

[16] Biber and Finegan (1986: 26) on the computer programs used for this research:

> "The lack of a large-scale dictionary combined with the large number of structural options which a particular grammatical construction can take limits the coverage of these programs. They were thus developed with a relatively modest goal: to capture 70-90% of the occurrences of a construction while avoiding any obvious skewing in any genre."

Constructions that were found to be problematic and had to be dropped from analysis include fronted 'that' clauses and initial prepositional phrases. (These could not be automatically recognized in transcriptions of spoken text.)

the lack of a proper methodology. Recent developments in the field of corpus linguistics, however, have provided the means to get access to data that up to then could only be obtained on a very small scale or not at all, while at the same time, various techniques for the manipulation and interpretation of data have been discovered and/or further developed (in this light the research by Biber and Finegan discussed above must be considered valuable). Therefore, it is our contention that future corpus-based research into linguistic variation potentially *does* have a substantial contribution to make with respect to the development of a variety theory. Crucially important in this respect is the way future research is set up and carried out. In our view a descriptive theory of linguistic variation should provide answers to at least the following three questions:

1. Under what extra-linguistic conditions is a particular variety used?

2. By what features is it described linguistically?

3. How do we describe the correspondences between linguistic and extra-linguistic categories?

Since, as Ellis and Ure (1969) point out, it is typical features rather than unique ones that are the subject of a variety study, a certain amount of prejudging is inevitable as part of the preliminaries to any variety study, in that texts will usually need to be grouped together (using an intuitive judgment of linguistic/extra-linguistic correspondences) to obtain a large enough initial corpus for study. Previous research may be surveyed to facilitate this prejudging. From what we have seen above, however, it is clear that although it may be wise to survey previous research in order to establish what linguistic and extra-linguistic features should be included in the next study, it also entails the danger not only of incorporating inadequately defined textual categories (while others may be overlooked), but also of restricting the scope of the research to those linguistic features that have previously been identified as possibly important without exploring the newly accessible data for other such features.

At this point it is evident that care must be taken both in the collection of data and the selection of features. The first presupposes the careful description of text categories in terms of their extra-linguistic variables (as Gregory suggested), and the compilation of a corpus for

the specific purpose of variety study. Such a corpus should meet at least the following criteria:

- the corpus should comprise a variety of samples which differ with respect to a number of extra-linguistic variables and should allow for the systematic varying of these variables;

- the individual samples of the corpus must be large enough to be representative of a particular variety; for the study of linguistic variation rather long samples are required. Experiences with a small subset of English have led us to believe that samples of 20,000 words each provide us with a much stronger foundation for quantitative studies than samples of 2,000 or 5,000 words. Results from previous studies (e.g. de Haan, 1989) indicate that, at least for rather frequent structures, a sample size of 20,000 words is sufficiently large in order to yield reliable information about their relative frequencies.

- the complete corpus must be large enough to make it possible to obtain ample quantitative data about the frequency of occurrence of various linguistic features as they occur in the different samples and allow for comparisons between samples.

Given such a corpus it will be possible to contrast texts or groups of texts and allocate those linguistic features that are characteristic of the texts that are being investigated to the corresponding extra-linguistic determinants. Co-occurrence patterns among linguistic features may be identified using sophisticated quantitative techniques such as the MF/MD approach Biber and Finegan propose. This approach will also make it possible to establish textual dimensions and to identify text types.

Little can be said about the total size of the corpus that will be needed for the study of linguistic variation, if one is to obtain statistically significant results. Much will depend on the number of variables involved which relates directly to the number of samples minimally required. Given the complexity of the matter and the fact that any research is bound by limited resources we would do well to proceed by studies on well-defined subsets of the language in individual research projects. One such project has recently been carried out at the University of Nijmegen, where a corpus of present-day British English was

compiled and (continues to be) investigated.[17]

Above some theoretical aspects have been discussed which relate to the study of linguistic variation on the basis of a corpus. Below attention will be given to the actual compilation of a corpus for the specific purpose of studying linguistic variation.

2.3.4 Designing a variety corpus

At Nijmegen University we took it as our objective to compile a corpus that would be representative of a well-defined subset of the English language. It soon became apparent that a 'well-defined subset' can only be established when reference is made to particularly clear external criteria. Thus we found that many criteria that had been provided in the literature and also criteria that were handed to us by experiences in other research projects,[18] could not be employed since they lacked a proper definition, or were too coarse or too refined. It was decided to restrict the corpus to a subset that could be described as 'written to be read', printed, educated, contemporary British English prose. Texts should be original British English publications, i.e. translations were not to be included. Likewise, texts that were of American English, Australian English, or Indian English origin were to be left out. By restricting the subset to prose we intended to exclude poetry. Similarly, the restriction to texts that were 'written to be read' led to the exclusion of plays, speeches, songs, etc., in other words, texts that are primarily meant to be spoken, recited, or sung rather than read. Since we intended to exclude private material such as personal letters, and also material that had only a limited distribution such as memos, we introduced the label

---

[17] The project referred to here is the second TOSCA project, which was funded by the Dutch Research Council for Advanced Research (NWO) from 1 March 1985 until 1 March 1989.

[18] Among these projects was the Nijmegen Computer Corpus Pilot Project. In this project a 130,000 word corpus was used, with samples of some 20,000 words each. Unfortunately, it appeared that although the samples represented various text categories, the samples could merely have an exemplifying function, pointing at obvious differences between some highly distinct text categories. As the subset of English represented in this corpus was too large considering the total size of the corpus (and the number of different samples), it was clear that the criteria on the basis of which the text categories could be distinguished were very coarse indeed.

'printed' which would ensure that the texts were both intended for and available to a wider public. Only texts in educated British English were to be selected, thus excluding non-educated and/or substandard English. Finally, in order to avoid the inclusion of archaic material we restricted ourselves to contemporary British English texts. 'Contemporary' was to some extent rather arbitrarily taken to be post 1975.[19] This was done for various reasons. Given the fact that the existing corpora were not suited to our objectives, we were forced to undertake the compilation of a new corpus. Since the research we have embarked upon will proceed for several years, we wanted our material to be as recent as possible - a desire shared by other researchers. The main British English corpus, the LOB Corpus, one may recall, takes 1961 as its sampling year. Another, minor[20] reason was our wish to avoid as much as possible any unpredictable influences of private time (the age of the author) on public time (the year of publication). Allowing for a briefer time-span was expected to minimize any differences in language use that might occur between the language use of, say, a 70-year-old in 1950 and that of a 20-year-old in 1987.

Once the subset had thus been established a number of initial text categories were defined. As we aimed at an acceptable minimum[21] of samples to represent each text category, we could only have a limited number of categories. This resulted in some instances in rather coarse categories such as religion and mythology (which makes up one category),[22] whereas other text categories that might well have been

---

[19] Actually, the average age of the texts at the time of selection (1985) was three years.

[20] We call this criterion 'minor' because so far there is no evidence that private time is a relevant variable in a study of linguistic variation.

[21] See below. The number of text categories was restricted for practical reasons. From a methodological point of view, however, it would be desirable to have not only a stronger representation of samples per text category, but also to distinguish between text categories that are more refined.

[22] The same goes for categories like 'psychology and psychiatry', 'sociology and anthropology', 'law and government', but also for categories such as 'biology', 'chemistry', 'health and medicine' and 'physics', where the labels define seemingly rather precise fields of discourse that cover, however, a large range of topics.

included were left out. Text categories that were included are

NON-FICTION

I Arts

NAUT	autobiography/biography
NEDU	education
NHIS	history
NLIN	language and linguistics
NLIT	literary criticism
NPHI	philosophy
NSOC	sociology and anthropology
NWOM	women's studies

II Science

NBIO	biology
NCHE	chemistry
NECO	economics
NGEO	geography
NMED	health and medicine
NPHY	physics
NPSY	psychology and psychiatry

III Miscellaneous

NGEN	non-fiction, general
NLAW	law and government
NMYS	mysticism and the occult
NPOL	politics
NREL	religion and mythology
NTRA	travel

FICTION

FCRI	crime and mystery
FHOR	horror
FHUM	humour
FNOV	general fiction, novel
FPSY	psychological novel
FROM	love and romance
FSFF	science fiction and fantasy
FSTO	general fiction, short story
FTHR	thriller and adventure

Sampling principles

Having established the text categories that were to be included in the corpus we then had to decide on the procedure that was to be followed in the sampling of the material. The following questions were raised:

1. What is to be the total size of the corpus?

2. What size should the individual samples be?

3. How many samples are required to represent a particular text category?

4. What samples should be selected within a certain text category?

5. What sampling procedure is to be followed within a selected source text?

We shall address each of these questions in turn.

As far as the total size of the corpus was concerned it was decided that the corpus should comprise at least one million words. A much larger corpus would hardly be feasible because of the time and money that would be required, while a smaller corpus was undesirable since it would force the subset to be further narrowed down if we were to maintain a certain minimum representation of samples per text category. Moreover, experiences with the 130,000 word Nijmegen Corpus had already demonstrated that such a corpus was far too small. With one million words the corpus would be comparable in size to the other major corpora, the American Brown Corpus and the British LOB Corpus.

In our view the size of the individual samples in the corpus had to be large since the corpus was to be employed in the study of linguistic variation. To have samples of 2,000 or 5,000 words as in other major corpora, would hardly yield representative samples of linguistic variation; rather, it would be more likely to show chance differences while neutralizing other, more significant differences. From experiences with the Nijmegen Corpus we knew that samples of 20,000 words each were sufficiently large in order to yield reliable information about the frequency of occurrence of most syntactic structures. A sample size of 20,000 words would yield samples that are large enough to be representative of a particular variety (see, for example, de Haan, 1984, and de Haan and van Hout, 1986).

As with the total size of the corpus and the number of text categories included, the number of samples per text category ideally should be as

large as possible. However, simple arithmetic shows that a corpus of one million words with samples of 20,000 words each allows for 50 samples to be selected. Given the text categories we wanted to be represented in the corpus the representation per text category was going to be rather poor which led to the decision to fix the number of words for the total corpus at one and a half million. The distribution of the 75 samples over the text categories selected was mainly based on intuitive judgment. Generally we aimed at a representation of a particular text category by at least two samples. For some categories, however, we chose to select a larger number of samples since these we felt were rather broad, e.g. 'general fiction', or 'biology'. A few of the miscellaneous categories -- because they merely had an exemplifying function -- were only represented by a single sample (e.g. 'mysticism and the occult', 'religion and mythology'). The distribution of the samples over the text categories then looks as follows:[23]

NON-FICTION (45)

I Arts (18)

NAUT	autobiography/biography	4
NEDU	education	2
NHIS	history	2
NLIN	language and linguistics	2
NLIT	literary criticism	2
NPHI	philosophy	2
NSOC	sociology and anthropology	2
NWOM	women's studies	2

Science (18)

NBIO	biology	4
NCHE	chemistry	2
NECO	economics	2
NGEO	geography	2
NMED	health and medicine	3
NPHY	physics	3
NPSY	psychology and psychiatry	2

III Miscellaneous (9)

NGEN	non-fiction, general	1
NLAW	law and government	2
NMYS	mysticism and the occult	1
NPOL	politics	2
NREL	religion and mythology	1
NTRA	travel	2

[23] The text categories are distinguished on the basis of the publishers' classification system.

FICTION (30)

FCRI	crime and mystery	3
FHOR	horror	2
FHUM	humour	3
FNOV	general fiction, novel	7
FPSY	psychological novel	2
FROM	love and romance	3
FSFF	science fiction and fantasy	3
FSTO	general fiction, short story	4
FTHR	thriller and adventure	3

The question what samples to select within a certain text category was rather central in the sampling procedure. As we observed above the corpus should make it possible to contrast texts or groups of texts and allocate those linguistic features that are characteristic of the texts that are being investigated to the corresponding extra-linguistic determinants. Therefore, samples could not be selected randomly. The selection of samples was obviously determined by a closed set of extra-linguistic variables. A number of the extra-linguistic variables that appeared in earlier studies had been accounted for in the definition of the subset. For example, taking the model suggested by Gregory (1967) such categories as temporal dialect, geographical dialect, social dialect and mode of discourse were reflected in the labels 'contemporary', 'British English', 'educated', and 'written to be read' respectively. Another extra-linguistic variable, the field of discourse, could be found in the distinction of text categories. One major extra-linguistic variable, however, had so far not been included: idiolect. Within this variable we distinguished between the author's identity, sex, age and origin. The author's origin was restricted to Britain in order to help to determine whether the language in a sample text could be looked upon as British (our first indicator of Britishness was original publication in England). The author's sex and age, although recorded in the documentation (Oostdijk, 1989), were considered minor variables and therefore not used as sampling criteria.[24] The author's identity, on the other hand, was considered of major importance, as it was expected that this might account for a linguistic variability that could not be attributed to any other extra-linguistic variables but that was due to personal style. In the selection of samples the author's identity was generally taken to be a

[24] However, only adult authors were selected.

free variable, i.e. it was not used to restrict the selection of samples. It was used, however, in some instances, to make it possible to investigate such issues as to what extent the personal style of an author affects the typicalness of a certain text category, or the (in)variability of personal style. Other extra-linguistic variables were text-length and distribution, both of which were determined by a practical motivation: since we wanted to be able to keep record of and control the extra-linguistic variables involved as much as possible, we restricted the selection to samples from novels that were written by a single author.[25] Newspapers, magazines, essays and articles were thus excluded. The distribution was to be national or international (as opposed to private or local).

Having decided how many samples to select from what texts one question remained: what sampling procedure should be followed with a selected source text? There was no scientific method that we knew of by which to select 20,000 words of running text. Rather than sampling from several instances in one particular source text we opted for a more or less random selection of the 20,000 words taking them from roughly the middle of the book.[26] Thus a sample was begun

- if a text had headed chapters or similar divisions, with the heading;
- if a text lacked chapter or similar divisions, with the first utterance following a blank line or similar spacing (as e.g. found after a diagram);

[25] In order to investigate what possible effect text-length could have on linguistic variation we included a few samples that only deviated with respect to this variable but otherwise conformed to the set of criteria.

[26] The decision to select samples from the middle of the book rather than the beginning or the end was to some extent rather arbitrary. However, it was decided to select the samples from the middle of the book rather than the beginning because it was felt that in this way any differences in the time writers take to introduce the various characters and so on would be neutralized. Then, for more practical reasons, we chose the middle section rather than the end because it would be difficult to determine at what place to start in order to get a 20,000 word sample running up to the very end of the book.

- if none of the above could be applied to a text, with the first utterance starting a new paragraph on the first page selected for sampling.

A sample ends with the utterance containing the 20,000th word. Headings are included in the text. Extra-textual material in the source (such as diagrams, maps, lists, bibliographies) is excluded, as are footnotes and references. Similarly long foreign quotations are excluded as well as all poetry.[27]

The processing of the material

The samples of text that were selected were keyed onto tape. In order to be able to retrace what the original text had looked like it was decided to have some form of coding of the text. For this purpose the coding that had been used in the processing of the LOB Corpus was adapted. Student-assistants were employed for the typing. These would type the text together with the codes required. Afterwards the text would be proofread and corrected, and a count would be made in order to yield a 20,000 word sample.[28] When this had been completed each sample of text was prefixed with a tag with the format

(XXXX TEXT nn)

where XXXX stands for a text classification code (e.g. NAUT) and nn for the rank number of the text in a particular text category. The XXXX code as well as the rank number of the text form the first part of the location code which precedes each line of the corpus text. A full location code occupies 13 positions: apart from the 6 positions taken up by the text classification code and the rank number, positions 7-10 are used to indicate the number of the page the text was on, and finally, positions 11-13 indicate the line number. Each sample of text is ended

---

[27] Material that was not included was generally replaced by tags. A full account of these tags and other coding symbols that were used in the processing of the text can be found in Oostdijk (1989).

[28] 20,000 words would be the minimum since a sample would end with the utterance that contained the 20,000th word.

with an end-of-subcorpus tag *#.

All information about a sample and its source text is contained in the manual that accompanies the corpus, including information on corrections or deviations from the original text.[29]

After the material had been selected and processed after the fashion described above, it was available for the compilation of word-frequency lists, concordances and such like. However, the insights that may be derived from this kind of word-based information are but few. Therefore, the use that was made of the raw corpus at this stage remained limited to extracting from it additional (word) information, which served to extend the available computer-readable dictionary. In the course of the project the raw corpus came to be used as a test-bed for the linguistic hypotheses that were laid down in the formal grammar. Parsing the corpus thus not only results in an analyzed corpus, it also provides valuable insights into the knowledge we have of the rules of grammar.

The care that has been taken in the compilation of the corpus and the recording of a great number of extra-linguistic features that characterize the texts should provide a firm basis on which to embark upon a study of linguistic variation, as was the primary objective of the second TOSCA project. The corpus, which serves as one of the two main language-specific input components to the TOSCA system that was described in chapter 1, constitutes a large potential of information that up to then was difficult to obtain. The key to this information in the approach adopted here lies in that other input component, the grammar.

While the ease with which the analyzed corpus can be studied for aspects of linguistic variation is very much a matter of the availability of a suitable database, the success with which this can be done depends on such factors as the adequacy of the analysis, the degree of consistency and the amount of detail in the analysis, all of which relate to the grammar. Therefore, in the next two chapters we discuss the design of the grammar (chapter 3) and some aspects of its implementation (chapter 4). Finally, in chapter 5 some analysis results are presented and an assessment is given of the role of the grammar in obtaining these.

---

[29] The manual in order is the *TOSCA Corpus -- Manual* (Oostdijk, 1989). A survey of the source texts that were used is given in Appendix E.

3 The Design of the Grammar

3.1 Introductory

In the approach to corpus analysis adopted by the Nijmegen group a central role is played by the (formal) grammar. It not only constitutes a useful tool in the production of databases containing detailed information about the linguistic structure of a great many sentences, it is also, as Aarts and van den Heuvel observe, "a powerful means of testing linguistic hypotheses which have been formalized in a grammar" (1982: 72). The use of a grammar as intermediary between linguist and parser sets the approach apart from others that adhere to hard-coded parsers.[1] The advantage of using grammar-based parsers rather than hard-coded ones lies in the fact that

> "The use of grammars makes it possible to cleanly separate the statement of the grammatical rules from the definition of the control mechanism that governs the application of these rules in the parsing process and from the maintenance of the records of recovered constituents. This facilitates both the correction and expansion of the grammar itself and the development of new parsing algorithms."
>
> (Halvorsen, 1988: 204)

Moreover, it is our contention that the grammar (also in a corpus linguistic environment where it is used as a tool) is linguistically interesting in its own right and should be accessible to other linguists, thus allowing for a discussion about the hypotheses embodied in it.

The present chapter is devoted to various aspects that relate to the general design of the grammar. First, however, since the role and the nature of the grammar appears to be a point of some controversy between the two major approaches to corpus linguistics, we start with a discussion of the role and the nature of the grammar in these two approaches, which we shall refer to as the 'Lancaster approach' and the 'Nijmegen approach'. Next an outline is given of the objectives that guide the construction of a grammar and determine what requirements should be

[1] Some corpora are being analyzed by means of hard-coded parsers. See for example Eeg-Olofsson and Svartvik (1984), and Francis and Kucera (1983).

met. In section 3.4 we discuss the formalism of Extended Affix Grammar that was used in the TOSCA project. Apart from giving a brief introduction (a more elaborate use of the formalism is exemplified in chapter 4), we consider the criteria that played a role in the selection of this formalism and discuss its merits with regard to these. We conclude this chapter by presenting a general outline of the structure of the grammar that was developed for the description and subsequent analysis of the material contained in the TOSCA Corpus.

3.2 The role and the nature of the grammar

Views on the role and the nature of the grammar in corpus linguistics have not always been the same, nor does there appear to be an international consensus with respect to these views. In this section we first address the latter issue, contrasting the role and the nature of the grammar in the Lancaster approach and the Nijmegen approach. After that, we consider the role of the (generative) grammar as it is employed in the Nijmegen approach.

As with the advances in computer technology in the seventies we gained access to the further potential of computer corpora, our focus shifted from the lexical level to the level of syntax and with it the concept of grammar came into play. The development of systems and their parsers that would make it possible to (automatically) syntactically analyze corpora of authentic language texts became a central theme in corpus linguistics. This is not to say that the objectives of the various approaches for the analysis of corpora are always the same. While more generally the primary objective is the creation of linguistic databases containing a wealth of information in the form of detailed analyses that may be used in the study of actual language use, others (cf. section 1.2) consider the analysis of corpora as a useful step in the process of developing a natural language processing system (as part of a larger application) that will be robust enough to process unrestricted input. It is, then, not surprising to find that these approaches take different views as to the role of the grammar. Indeed, the role of the grammar has been and continues to be the major point of difference between two of the main corpus-based approaches, the Lancaster approach on the one hand, and the Nijmegen approach on the other. Meanwhile, the implications for the nature of the grammar are rather substantial. While the mainstream of computational linguistics, including the Nijmegen corpus-based

approach, make use of some form of generative grammar in order to analyze their data, the Lancaster approach rejects this use, claiming that only non-generative, probabilistic means hold the potential to deal with all of the data adequately.

The stand taken by the Lancaster group towards the use of generative grammar to parse natural language is most strongly expressed in one of the statements that Sampson on various occasions has put forward on the subject. Sampson (1987a: 20) observes that

> "Much of the difficulty of designing generative-grammar-based processing systems lies in the fact that, if a grammar is complex, it is hard to design an algorithm which succeeds in locating the consequences of the grammar for particular strings. If the activity of revising a generative grammar in response to recalcitrant authentic examples were ever to terminate in a perfectly leak-free grammar, that grammar would surely be massively more complicated than any extant grammar, and would thus pose correspondingly massive problems with respect to incorporation into a system of automatic analysis.
>
> Accordingly, the idea of basing automatic language-processing on generative grammars of any category seems to me a dead end."

The key issue in Sampson's argument is the impossibility of constructing what he refers to as a leak-free grammar. Such a grammar should describe all and only the grammatically well-formed sentences of the language. Reviewing the systems that "embody the generative-linguistics concept" he observes that

> "... such systems incorporate some type of generative grammar (whether in the form of an ATN, a GPSG, or whatever) which defines the class of 'all and only' the inputs which the system is expected to parse, and 'deviant' sentences are ignored, or handled by some more or less *ad hoc* subsidiary mechanisms. Researchers working on such systems seem commonly to run them over invented examples only."
>
> (Sampson, 1987b: 219)

Sampson then continues by pointing out that authentic natural language material as contained in a corpus demonstrates that the grammatical/ungrammatical distinction is not at all as clearcut as to grant the validity of the all-and-only hypothesis. This, in convergence with the vast diversity of constructions that are found to occur, leads him to conclude that the formulation of an adequate generative grammar is "doomed to failure". Sampson -- advocating the Lancaster non-generative, probabi-

listic approach -- submits "objective evidence" in the form of a paper reporting on research bearing on the subject, arguing that the choice between the alternative approaches has always been made in a subjective fashion: "some researchers are impressed by the power of generative-grammar formalisms, others are impressed by the messiness of the data" (1987: 220). Briscoe (1990), however, disproves the claims that Sampson derives from this research, finding the methodology Sampson adopted at fault:

> "Sampson's result is suggested by his *analysis* of this data, not the data itself."
>
> (1990: 57)

Briscoe argues that

> "Whilst it seems likely that 'all grammars leak' slightly, one clear problem with Sampson's argument is that his evidence only bears on one particular and implausible generative grammar, rather than on the paradigm as a whole. It may well be that the generalisations which can be expressed in terms of a phrase-structure grammar employing a finite set of (nearly) atomic categories are not those appropriate to elegant description of natural language syntax (Chomsky, 1957, Gazdar et al., 1985)."
>
> (1990: 47)

To illustrate this point Briscoe reports on research carried out by Taylor et al. (1989), the results of which he observes

> "... demonstrate quite clearly that a feature-based unification grammar employing a recursive and 'deeper' style of analysis captures the relevant generalisations more efficiently than the analysis and implicit formalism employed by Sampson (1987a)."
>
> (1990: 57)

Briscoe admits that with the application of the grammar in the research carried out by Taylor et al. "there is good reason to believe that 'all grammars leak', slightly." However, failure of analysis appears due to oversights in the grammar rather than "syntactically mysterious" variation.

Most of Briscoe's reply to Sampson is about the alleged inadequacy of generative grammars for the purpose of natural language processing where the obscurity of the grammatical/ungrammatical distinction and the vast diversity of structures make it futile, according to Sampson, to use these grammars. An issue that is not taken up by Briscoe is the fact that throughout his argumentation against the use of generative grammars Sampson maintains the view that a leak-free generative grammar describes all and only the well-formed sentences of the language. Although this view may be upheld when a theory of the language is under construction, or when application of the grammar for the purpose of generating sentences is intended, the question arises whether this should also be the case in corpus analysis. Since in current corpus linguistic practice the creation of databases is given priority over other aims, must not a grammar that is 'more permissive' in the sense that it overgenerates and, when put to analysis, assigns the appropriate structure to all the sentences of a language, be considered adequate, in spite of the fact that it may not be able to distinguish at all times between grammatical and ungrammatical sentences? Given the present scope of our grammars (see below) and the analysis pursued it appears impossible to view this matter any differently. This is not to deny that ideally (eventually) the objective should be the development of a grammar that does describe *all and only* the well-formed sentences of the language. However, at present it is, in our opinion, a realistic and respectable objective to construct grammars that are able to assign a description to *all* sentences that occur in a corpus of texts, assigning to them a structure that is intuitively correct and reflects the state of the art in English descriptive linguistics.

Allowing for generative grammars of the kind described above (i.e. grammars that are more permissive), little remains of Sampson's argumentation against their use. Seen in this light, Atwell's plea for the use of probabilistic methods, such as embodied in for instance the constituent-likelihood grammar used in the Lancaster approach, loses its significance. On the use of constituent-likelihood grammar set against the use of generative grammar Atwell observes

> "A 'constituent-likelihood grammar' does not define the set of sentences (symbol-strings) which constitute a language, to the exclusion of all other possible sequences of symbols. No linguist (or team of linguists) has yet come up with an adequate generative grammar which truly generates all possible English sentences (and does not generate any non-English sentence); since the syntax of a natural language such as English is extremely complex, large corpora of texts will continue to

> throw up sentences which are not dealt with adequately. Furthermore, in many applications of syntactic analysis we must assume that the input may be noisy, that is, the text may contain errors; again, noisy text will not be dealt with adequately by a generative grammar."
>
> (Atwell, 1987: 57)

Two observations must be made at this point. First, there is no empirical evidence whatsoever indicating that the syntax of a natural language such as English is too complex to be described in terms of a comprehensive (generative) grammar. Second, the use of a probabilistic grammar can only lead to descriptions and analyses that lack a great deal of detail and fail to capture linguistically relevant generalizations. Thus, in developing a probabilistic grammar it is common practice to refrain from including a great amount of detail since this would have an immediate effect on the probability assigned to various alternatives. With the increase of the amount of detail, the number of alternatives is likely to increase as well and as a consequence the average probability of the alternatives decreases. With a parser that produces only the most probable analysis at any one time, the chance that the same analysis is obtained in subsequent parses of one and the same string is reduced considerably when a great many alternatives may be applied with a similar likelihood. Moreover, if this parser appears to be inconsistent in the parsing of a single string, it will fail to assign similar structures to similar strings when parsing a corpus, i.e. it will fail to capture linguistically relevant generalizations. Here, again, the fact that the Lancaster approach is application-oriented rather than purely linguistically oriented may account for the opinion put forward by Atwell. Moreover, one may seriously want to question whether in corpus analysis for purely linguistic purposes we should actually pursue the analysis of erroneous input.

As a conclusion to his introduction to constituent-likelihood grammar Atwell puts forward the following thought:

> "The constituent-likelihood approach to grammatical description even opens up the intriguing possibility of automatically extracting a grammar from *raw, unanalysed* text corpora ..."
>
> (Atwell, 1987: 65)

If one takes the view of the grammar described in the introductory section to this chapter, i.e. as a means for producing databases *and* as a means for testing the linguistic hypotheses embodied in it, the outlook Atwell presents holds no promise, since it restricts the conception of the grammar in every possible way. The kind of grammar Atwell appears to advocate is a grammar of just the utterances contained in the corpus rather than a grammar of the language of which the utterances found in the corpus are but instances.

The generative grammar in the Nijmegen approach has always had the double role described above, even in the Computer Corpus Pilot Project (CCPP) which preceded the TOSCA projects. Experiences in the CCPP -- the drawbacks of an independent manual tagging phase preceding the automatic analysis on the one hand, and the positive judgment of the use of a formal grammar on the other -- influenced the design of the analysis system to be developed in the first of the TOSCA projects. Acting upon the desire to eliminate the need for manual tagging, the grammar was given a more prominent role than before. Commenting on the design of the system and the role of the grammar Aarts and van den Heuvel observe that

> "Since the formal grammar cannot be expected to describe fully the semantics of a natural language, we shall have to make an appeal to the user's knowledge of the semantics of the corpus language. The system should therefore be set up in such a way that it is easy for the linguist to intervene in the analyzing process; he should be able to select one out of many parses as the semantically most plausible one or to make a correction if analysis fails. Intervention is also needed to enable the linguist to judge the adequacy of his grammar; where it proves to be inadequate he can decide either to make changes in the grammar or to provide additional information to the system."
>
> (Aarts and van den Heuvel, 1982: 73)

Writing a grammar is conceived of as a (basically) cyclic process. The first version of the grammar is written on the basis of the linguist's intuitions and his knowledge of the structure of the corpus language as well as any information that may be found in handbooks of grammar. Testing the grammar on a collection of test sentences brings to light any flaws and/or lacunae in the grammar. It is then extended and/or revised and the latest version is tested on further testing material. This process is continued "until the grammar has reached the desired level of

completeness" (Aarts and van den Heuvel, 1985: 309). Then the grammar may be put to use in the analysis of the corpus. Note that with respect to this point again there exists a major difference between the Lancaster and the Nijmegen approaches. While Sampson (and Atwell for that matter) denounces the use of intuitions -- "it is clear that an automatic NL-analysis system ... must be based on empirical investigation of which constructions do occur in practice, rather than on linguists' intuitions about what constructions might be expected to occur" (Sampson, 1987b: 220) -- in the Nijmegen approach intuitions and empirical data are found to interact.

3.3 Objectives and requirements

As was observed above, the aim of corpus linguistics is to gain (further) insight into language, its use and variation in that use. It does so by studying authentic natural language text as contained in various corpora. The analysis of corpora serves a double role: (1) it yields databases containing detailed information that may be used by linguists, including linguists from other linguistic subdisciplines, in investigating particular phenomena; (2) it allows for the testing of linguistic hypotheses that have been formulated in terms of the formal grammar. For the time being, however, the creation of databases is given priority over optimizing our descriptive model, i.e. adapting and extending it in the light of newly acquired insights. The reason for this is that, as Aarts and van den Heuvel argue, "this is the most urgent task" (1985: 310). Even to date corpora that have undergone a (detailed) syntactic analysis are but few. Making analyzed corpora generally available therefore continues to be the primary objective in corpus linguistics. As Aarts and van den Heuvel point out

> "Not until sufficiently large and reliable databases have become available will it be possible to separate the two purposes and to experiment with grammars that extend linguistic theory and descriptive domain."
>
> (1985: 333)

Meanwhile, as the creation of databases is given priority over the optimization of our linguistic descriptive model and grammars are being constructed for the purpose of analyzing corpora, the linguist's freedom is restricted by what is familiar and traditional in linguistic description.

This is not to say, however, that traditional descriptive practice is maintained in each and every instance. Although tradition and familiarity may guide the choice of what general descriptive framework is adopted, lacunae and inconsistencies encountered in traditional descriptive grammars give cause to devising alternative descriptive strategies (see also chapter 4).

In addition to the choice of a general descriptive framework, the design of the grammar further comprises making decisions as to the scope of the grammar and its coverage. Each of these aspects is discussed below. Attention is also given to the points of conflict of interest that arise when such issues as autonomy and efficiency of the parser are brought into play.

scope

While corpus analysis eventually aims at the full, detailed analysis of running text, i.e. including aspects of a semantic and pragmatic nature, current corpus linguistic practice consists in the morpho-syntactic analysis of individual utterances. As yet, formalized descriptions of unrestricted input do not extend to include the level of semantics and/or pragmatics, nor do they comprise a description of text structure.

The current restricted scope of the grammar affects the analysis process directly since it yields little autonomy to the parser. Since the grammar describes merely the level of syntax it generates a certain amount of ambiguity. Consequently, interventions are needed to effect the desired disambiguation by supplying additional information in such areas as semantics, text structure, general knowledge about the world. It is our experience that apart from being a time-consuming task, interventions are often made at the expense of the much-valued consistency of analysis.

coverage

As the value of the corpus as a linguistic database depends in part on the 'degree' of analysis[2] -- i.e., both the amount of detail as well as the percentage of utterances that have received an analysis -- the question

[2] Other factors are, for instance, the scope and consistency of the analysis, and the composition of the corpus.

arises whether the objective to provide each string (utterance) in the corpus with an adequate structural analysis entails the analysis of *all* strings or of *all grammatical* strings.[3] The answer, it would appear, is neither. The analysis of all strings would presume that they could all be judged acceptable. Although it would be very convenient to postulate the acceptability of all strings in the corpus, this is counterintuitive since we do come across strings that the writer intended to be ungrammatical. As Aarts (1991) points out, "every corpus will contain sentences that the writer wanted to be ungrammatical (in the widest sense of the word) by either violating the rules of the grammar, or by using elements from a substandard variety of the language." Similarly, the analysis of all strings would also entail the analysis of material that is overtly erroneous. Restricting the analysis to only the grammatical strings on the other hand, one would fail to incorporate those strings that, although 'ungrammatical', are found to be perfectly acceptable.

What then must the grammar describe? This constitutes what Aarts in a recent paper[4] has termed "the linguist's dilemma". As we gain more experience in the formal description and subsequent analysis of corpus material this issue becomes more and more central to corpus linguistic practice. Where does one stop incorporating the description of (possibly highly irregular) structures? Practice so far has amounted to incorporating, in principle, the descriptions of structures that were taken to be 'current', i.e., structures that were judged to be generally accepted and observed to be relatively frequent, while they should be amenable to the descriptive framework employed. It must be admitted that this currency criterion still heavily relies on intuitive judgments, for it is as yet not quite clear how the criterion relates to such notions as 'grammatical' and 'acceptable'. Occasionally, structures that were judged to be current were all the same not accounted for in the grammar; they include those structures whose description would introduce a large amount of undesired ambiguity in the analysis of other, highly frequent ones.

One aspect that has received little attention so far is the degree of efficiency that is achieved in the analysis. Some corpus linguists have

[3] Of course if by 'grammatical' we mean 'defined by the grammar (underlying the parser)' the two could mean the same. This interpretation of the term is, however, not intended here.

[4] Paper presented at the tenth ICAME Conference held at Bergen (Norway), 1989.

argued that this is "linguistically irrelevant" (cf. van den Heuvel, 1987). Strictly speaking they are right. On the other hand, in the actual analysis of a corpus we are bound by limited resources and therefore efficiency *does* matter. It is undeniably so that there is a certain amount of tension between the linguistic and computational interests when it comes to the question how things are best described.

3.4 The formalism of Extended Affix Grammar

For reasons that were set out and discussed above (see section 3.1), we opted for the use of grammar-based parsers in Nijmegen rather than hard-coded ones. Experiences in the Computer Corpus Pilot Project where a context-free grammar (CFG) had been used to parse material that had been tagged manually, had demonstrated that the context-free grammar -- although very efficient in parsing -- tended to become unwieldy. Not only did the sheer size of such a grammar appear problematic for the machine, it also was found to be not very attractive from a linguistic point of view. The fact that context-free grammar is not very economical was experienced as a drawback, especially since this caused apparent generalizations to be easily obscured. Therefore, in the TOSCA project, which aimed at the analysis of unrestricted input, another formalism was employed, that of Extended Affix Grammar (EAG).

EAG was developed in computer science for the definition of artificial languages. Since then it has been applied and adapted to the description of natural languages, such as English. Here we include a brief introduction to EAG, but refrain from giving a formal definition. Instead, the reader is referred to Koster (1971), Kühling (1978), Meijer (1986) and Watt (1974).

EAG is a type of two-level grammar. In fact an EAG consists of two context-free grammars that are rolled into one. The context-free grammar that constitutes the first level is an ordinary CFG, i.e. a set of terminals and non-terminals and a set of rewrite rules. To the non-terminals of this grammar so-called affixes can be attached. The use of these affixes gives the grammar minimally the power of a context-sensitive grammar. The affix level constitutes the second level of the grammar. The affixes are also defined by a CFG, the meta-grammar. The term 'affix' here should not be understood in a linguistic sense; rather, an affix can be looked upon as a kind of parameter that is used

to transfer information from one point in the grammar to another, or as a means to impose restrictions on certain (first-level) rules of the grammar. Such restrictions can remain more or less implicit, or they can be stated explicitly in so-called predicates.

Like a context-free grammar an EAG can be automatically converted to an analysis program, i.e. a parser, by means of a parser generator. One of the reasons for opting for EAG was the availability of such a parser generator for EAGs at Nijmegen University, where an EAG parser generator had been developed at the Computer Science Department. Research in this area continues and comprises the investigation of possible optimizations and extensions. For example, more recent developments include the development and implementation of Affix Grammar over Finite Lattices (AGFL; van Zwol, 1990; Koster, 1991). In the formalism of AGFL the affix domains are restricted to finite sets of values. Therefore the parsing process can run more efficiently than in the case of EAG where unrestricted affix domains are assumed. Close cooperation with the Computer Science Department has given us the advantage of being able to benefit from a great deal of expertise with respect to parsing methods and related matters.

The construction of an EAG: The context-free level

Below we present a simple CFG which will form the starting point in our construction of an EAG. The notational conventions deviate slightly from the ones usually employed in formal linguistics. For instance, a colon is to be interpreted as the instruction 'rewrite as'; by means of a semi-colon alternatives are distinguished, whereas a comma is used to separate the members of a rule. Each rule is delimited by a period. The rule specifying the start symbol is distinct from others in that it does not contain a colon, i.e. in this rule no rewriting takes place. Terminal symbols are enclosed in quotation marks. Non-terminal symbols may consist of strings of letters and digits, including optional blanks.[5] Our CFG looks as follows:

[5] Although not part of the formalism, we find it convenient to use small letters for terminal and capital letters for non-terminal symbols at the context-free level, with the exception of non-terminals that occur in predicate rules. At the affix level small letters are used for non-terminal symbols, capital letters for terminal ones.

```
( 1) SENTENCE.
( 2) SENTENCE : NP, VP.
( 3) NP       : DET, NOUN;
                PRONOUN.
( 4) NOUN     : BASE;
                BASE, PLURAL SUFFIX.
( 5) VP       : REGULAR VERB;
                REGULAR VERB, PERSON MARKER.

( 6) PRONOUN  : " I"; " you"; " he"; " she"; " it"; " we"; " they".
( 7) DET      : " the".
( 8) BASE     : " aeroplane"; " passenger"; " building"; " plan".

( 9) PLURAL SUFFIX    : "s".
(10) REGULAR VERB     : " land"; " arrive"; " collapse"; " remain".
(11) PERSON MARKER    : "s".
```

Note that the grammar describes both acceptable and unacceptable strings, e.g.

```
  the aeroplane lands
  the passengers arrive
* I remains
* he collapse
```

Supplementation with affixes

We may now supplement this CFG with so-called *affixes*. These can be attached (or: affixed) to any non-terminal of the context-free level of the grammar. The affixes are enclosed in brackets and separated by commas. It is left to the writer of the grammar to decide what kind and what number of affixes are to be introduced at what point in the grammar, although he is bound by certain conditions with respect to the number of affixes and their name-giving. For instance, one of the conditions under which the affixes operate is that within one and the same rule identical affixes denote identical values.

In order to describe the concord relation between the subject and the verb of a sentence, we introduce the affixes 'number' and 'person'. Consequently, rule (2) is extended to

(12) SENTENCE : NP (number, person),
VP (number, person).

By means of rule (12) we express the fact that a sentence may consist of an NP with a specific number and person, followed by a VP with the same number and person.

Whereas the above condition takes care of the relation between affixes within one and the same rule, we may formulate a second condition to establish the relation between affixes in different rules. This condition is two-fold:

1. the same non-terminal must have the same number of affixes at every occurrence;

2. corresponding affixes in defining and applied occurrences (see below) of a non-terminal must have identical values.

The first part of this condition can be illustrated by means of the following rules:

(13) NP (number, person) : PRONOUN (number, person).
(14) NP (number, "3RD") : DET, NOUN (number).

In both rules the non-terminal NP is supplemented with the affixes 'number' and 'person'.[6] Now if we compare the occurrence of NP in rule (12) to that in rules (13) and (14), we can say that in rules (13) and (14) the notion NP is defined. In rule (12) on the other hand, it is the notion of sentence that is defined, applying the NP definition of rule (13) or (14). Thus a distinction can be made between the non-terminal NP in rule (12) which we call an *applied* non-terminal, and the non-terminal NPs in rules (13) and (14) which are called *defining* non-terminals. The second part of our condition then amounts to saying that given e.g. rules (12) and (14), a relation holds so that 'number' in rule (12) is rewritten as the 'number' of rule (14), and 'person' in rule (12) is rewritten as "3RD" (rule 14).

[6] Note that these rules cannot be conflated since their left-hand sides differ on the affix level.

Meta-rules

Above we formulated two conditions under which affixes operate. We showed how relations between affixes could be established when these occur in the same rule and also when they occur in different rules. In order to ensure that the affixes will be assigned the correct values no further conditions are needed. We may, however, introduce so-called *meta-rules*. The effect on the grammar is two-fold:

- first, the use of meta-rules reduces the bulkiness of the grammar which we would otherwise have;
- secondly, the readability of the grammar is improved.

Meta-rules specify the possible values -- terminal or non-terminal -- that a particular affix can take. This specification consists of a number of alternatives which can be (terminal) literal affix values, or (non-terminal) affix names or expressions. An example of a meta-rule is:

(15) number :: "SING"; "PLU".

By means of this rule we express the fact that the affix 'number' can take the (literal) value "SING" or "PLU". Meta-rules are no different from any other rewrite rules. The only difference lies in the notation: instead of a colon we now use a double colon, and instead of a comma we use a plus sign to separate members. Further notational conventions remain the same. Note that the meta-rules together constitute a CFG.

Affix expressions

An affix position may be occupied by more than one affix name or literal affix value. When this is the case we speak of *affix expressions* rather than affix variables or literal affix values. An affix expression is a sequence of affix names and/or values concatenated or separated by plus signs. Affix expressions may occur in any affix position.

Affix expressions may be used if two or more affixes are in some way dependent on each other or relate to different aspects of one and the same phenomenon. For example, the affixes 'number' and 'person'

that were introduced in order to describe the concord relation that holds between subject and verb each describe one particular aspect of this relation. Rather than scattering the information over various affix positions, we can concentrate it on one single affix position by means of an affix expression. Rule (16) then replaces rule (12):

```
(16) SENTENCE : NP (number + person),
                VP (number + person).
```

We are now in a position to decide that in the rule for sentence we do not want to concern ourselves with the aspects of concord. If this is the case, we can replace the affix expression 'number + person' by the affix variable 'concord'. As a result we get

```
(17) SENTENCE : NP (concord),
                VP(concord).
```

In the meta-rules 'concord' must be defined as

```
(18) concord :: number + person.
```

Two further meta-rules define the affixes 'number' and 'person':

```
(19) number::   "SING"; "PLU".
(20) person ::  "1ST"; "2ND"; "3RD".
```

In the grammar the affix 'concord' will be used whenever we do not wish to concern ourselves with any details relating to the respective values of the constituting affixes. Only where relevant is the affix 'concord' analyzed again into the affixes 'number' and 'person'. Thus, whereas we would change rule (13) into rule (21):

```
(13) NP (number, person)  :  PRONOUN (number, person).
(21) NP (concord)         :  PRONOUN (concord).
```

we would retain the affixes 'number' and 'person' in a rule like (14), but now 'number' and 'person' form one affix expression:

(14) NP (number, "3RD") : DET, NOUN (number).
(22) NP (number + "3RD") : DET, NOUN (number).

Predicates

Apart from meta-rules there is one more device which we may use, namely so-called *predicates*. Predicates are rewrite rules with empty right hand sides that are used to impose restrictions on or to effect the generation or analysis of a particular affix value elsewhere in the grammar. For example, if we were to describe the 'additive' coordination, i.e. coordination by means of the coordinator *and*, of two NPs in subject position, we could formulate the following rule, in which we call upon a predicate ('additive person ...') in order to yield the correct value for 'person' for the subject:

```
(23) SUBJECT (person):
        NP (person1),
          COORDINATOR ("ADD"),
            NP (person2),
              additive person (person1, person2, person).
```

where the meta-rules for 'person', 'person1' and 'person2' may be assumed to be

```
(24) person   ::   "1ST"; 2nd or 3rd.
(25) 2nd or 3rd ::   "2ND"; "3RD".
(26) person1  ::   person.
(27) person2  ::   person.
```

By means of the coordination-rule instances like

the man and I
you and me
he and his brother

can be described. With each NP some value for person is associated. For example, "the man" will have the value "3RD" for person, while "I" has the value "1ST". The value for 'person' for each of the NPs involved in the coordination may -- but need not -- differ, hence the distinction that is being made by means of the indexed 'person1' and 'person2'. The predicate 'additive person (person1, person2, person)' is introduced to effect the value for 'person' for the coordination of NPs, in other words, the value for 'person' that should be associated with the subject. The predicate takes the values for 'person' that were associated with the first NP ('person1') and the second NP ('person2') and yields the value for 'person' for the coordination. In the case of additive coordination the following set of predicate rules applies:

(28) additive person (person, person, person): .
(29) additive person ("1ST", 2nd or 3rd, "1ST"): .
(30) additive person (2nd or 3rd, "1ST", "1ST"): .
(31) additive person ("2ND", "3RD", "2ND"): .
(32) additive person ("3RD", "2ND", "2ND"): .

By means of this set of rules we can effect the correct value for 'person'. Thus the value for 'person' will be "1ST", in case both 'person 1' and 'person 2' have the value "1ST", or in case 'person 1' has the value "1ST" and 'person 2' the value '2nd or 3rd', or in case 'person 1' has the value '2nd or 3rd' and 'person 2' the value "1ST".

3.5 The structure of the grammar

In the previous sections the term 'grammar' has been used in various contexts, signifying different things (cf. Greenbaum, 1988: 20ff). For instance, it was used -- in accordance with the more common definition of the word -- in the sense of *an ideally complete description* of a language, but also in the much more restricted sense of *the description of the syntax* of a particular language. While referring to context-free grammar and extended affix grammar, the word 'grammar' was used both in a technical sense, meaning grammar *formalism*, and also in the sense of a *formal description* (not necessarily linguistic). In the present context two further senses may be associated with the word 'grammar', namely (1) *the formalized (linguistic) description* of a language and (2)

the formalized (linguistic) description of the syntax of a particular language. It is the structure of the grammar in the latter sense of the word that forms the topic of the current section.

In section 3.2 we noted that as current corpus linguistic practice is primarily concerned with the creation of databases containing analyzed corpora, we are restricted in our grammars by tradition and what is familiar in linguistic description. Traditional linguistic descriptions of English syntax, however, lack any degree of formalization and (worse) some even fail to (properly) explicitly define their linguistic descriptive terminology. A grammar such as Quirk et al. (1985, 1972) is a typical example of a grammar in this descriptive tradition. Although it has an impressively wide coverage, incorporating all sorts of structures and phenomena commonly found in English, much is left implicit and descriptions are found to be inconsistent from time to time. Writing a grammar for purposes of corpus analysis therefore constitutes a great deal more than merely formalizing a given traditional description.

The construction of a formal grammar is preceded by such preliminaries as determining *what structures* to incorporate in the grammar and -- equally important -- *how* to incorporate them. The latter point amounts to establishing what descriptive system should be used, and also what formalism is most suited for the purpose. While the selection of the formalism is generally motivated by arguments relating to its power and the possibility of automatic parser generation[7], the choice of the descriptive system is governed in part by linguistic tradition and in part by the views held by the grammarian. Formalization forces us to be explicit and exhaustive in our descriptions. Moreover, as a formal grammar is converted into a parser and used for analysis, consistency in the analyses is guaranteed. Formalization does, however, not warrant the consistency of the linguistic description. Therefore, it remains the grammarian's responsibility to design a descriptive system that is inherently consistent and to implement it as such.

Since the approach underlying handbooks of English grammar such as those by Quirk et al. (1985, 1972) is subscribed to by a great many linguists, it was decided to base our descriptive system on that put forward by Aarts and Aarts (1982) and make whatever amendments were necessary. The approach Aarts and Aarts present is basically similar to that which we find in Quirk et al., but unlike the latter they have a

[7] Other arguments include the expertise available and the efficiency. See also section 3.4.

much more rigid descriptive system. In this system a structure is assumed which is based on immediate constituency and which presents the rank scale. Constituents are labelled for their function and category. Thus the labelling holds information both about the syntactic characteristics of a single descriptive unit which it shares with other units of the same kind, and about its role in a larger linguistic structure (cf. Aarts and Aarts, 1982: 4-14). Although Aarts and Aarts at this point are not very specific about what such roles constitute, we find that the notion of function is best defined as the syntactico-semantic role of a constituent. In a grammar like ours in which the scope is restricted to the level of syntax and which cannot appeal to intuitions about any semantic relations that hold between constituents, such a conception of the notion of function can no longer be maintained. Moreover, the multilayered structure that was introduced throughout the grammar for the handling of coordination so as not to have to postulate ellipsis in almost every instance (as is the case with surface structure descriptions as found in Quirk et al., and Aarts and Aarts), associates the notion of constituent with (parts of) structures that are not traditionally looked upon as constituents. For example, in describing the coordination in sentences like

(33) The shopkeeper gave my friend an apple and me an orange.
(34) From her hotel the mother sent her daughter a letter and her son a postcard.

we do not postulate the coordination of two sentences with ellipsis of the subject and the verb in the second conjoin; rather, the indirect object and the direct object are considered to form one constituent, i.e. a node 'ditransitive complement' is postulated, which has a coordinated realization (see also section 4.2). Consequently, while a function may be conceived of as the syntactic role a constituent plays in a higher order constituent, it appears impossible to associate any semantic role with it. The function-category dichotomy has been retained, however, since -- even with this revised conception of the notion of function -- functions still prove to be useful in the grammar. Not only do they constitute a means of indicating the syntactic relations that hold between constituents, they can also be used as variables over categories. Thus valuable generalizations about the distribution of categories (e.g., what categories can realize the function of subject) can be expressed.

Apart from restricting the notion of function from the syntactico-semantic role of a constituent to its syntactic role, we also assessed the given rank scale. Aarts and Aarts distinguish between morphemes, words, phrases and clauses/sentences.[8] A word may consist of a single lexical item or a combination of lexical items. In the latter case a word is referred to as a multi-word. Words may be grouped together in word classes. Occasionally a word belongs to different word classes, in which case it has multiple class-membership and can be said to be lexically ambiguous. "A *phrase*", according to Aarts and Aarts' definition, "is a constituent which can be identified on the basis of the word class membership of at least one of its constituent words, whereas a *sentence* (or *clause*) is identifiable on the basis of the relations holding among its immediate constituents" (1982: 60). Aarts and Aarts distinguish five types of phrase: noun phrase, adjective phrase, adverb phrase, verb phrase and prepositional phrase. All of these, with the exception of the prepositional phrase which is an exocentric construction, are endocentric constructions in which there is one obligatory (functional) constituent, the head, while other constituents are optional. Finally, sentences constitute the largest units in the descriptive system. When the implications of the given rank scale were considered, it soon became apparent that in describing such phenomena as coordination and transposition[9] the descriptive units (the word, phrase and sentence/clause) caused the descriptive system to be too rigid, in the sense that it did not allow for constituents that were parts of phrases or sentences/clauses. In other words, whenever a constituent was larger than a word and smaller than a phrase it could only be considered to be an elliptic phrase; a constituent larger than a phrase but not a full sentence/clause was considered to be a clause. Having to postulate ellipsis in all instances of coordination,

[8] The distinction between the sentence and the clause according to Aarts and Aarts is as follows: the sentence is "regarded as the largest unit of grammatical description since it does not function in the structure of a unit higher than itself" (1982: 79). On the other hand, "sentences that are embedded in the structure of other sentences or in the structure of phrases are called *clauses*" (1982: 80).

[9] The notion of transposition that is introduced here is a collective term which includes a wide variety of phenomena which have in common the displacement (i.e. fronting/pre-positioning or postponement) of a constituent in the widest sense of the word. For instance, transposition includes the postponement of modifiers but also topicalization and inversion. See also section 5.3.

transposition, etc. that do not involve full phrases or sentences/clauses was considered quite unsatisfactory, since it meant a departure from the actual surface structure. Therefore, in our adaptation of the Aarts and Aarts framework we extended the notion of phrase so as to yield a more flexible system that would allow for a description on the basis of the elements that are actually found to be present in the surface structure. A phrase then can be defined as a higher order constituent which consists of one or more immediate (functional) constituents.

In the grammar we adopted a multi-layered structure which allows for almost any two adjacent constituents to combine into a higher order constituent.[10] As a consequence the grammar is rather flexible when it comes to associating a structural analysis with a particular input string. Where in principle this is considered to be an advantage since the grammar can thus handle a vast variety of structural variants, a number of restrictions must be introduced to avoid any spurious ambiguity that might otherwise result from it. Here the strict alternation between functional and categorial constituents comes into play. The combined use of functions and categories serves as a control mechanism that is used to restrict generalized descriptions, such as the description of coordination that is discussed in section 4.2. Rather than allowing for any two constituents, whether functional or categorial, to be coordinated, we assume coordination to take place between categorial constituents under dominance of one and the same function node. Thus we avoid the inconsistency of analysis and resulting ambiguity that we find in Quirk et al.'s description of coordination. For example, Quirk's analysis of the coordination we find in (32) is four-fold ambiguous.

(32) John and Harry are brothers.

One analysis consists in taking *John and Harry* to be a coordination of proper nouns. A second analysis places the coordination at the level of the NP HEAD function, i.e. it postulates the coordination of two NP HEADs. A third analysis is one in which two NPs are assumed to be coordinated. Finally, a fourth analysis interprets the coordination as a coordination of subjects. Restricting coordination to the coordination of categorial constituents under dominance of one and the same function node yields a consistent analysis for coordinations (they are all coordi-

[10] See also sections 4.2 and 4.3.

nations of categories), while at the same time it reduces the ambiguity by half.

In writing the grammar we opted for a modular structure. This makes it possible for the grammarian to deal with one structure, or set of structures, at a time. In writing separate modules for subsets of (related) structures, he can concentrate on one particular (type of) structure before continuing with the next one. Thus what would otherwise undoubtedly be a fairly complex task is divided into convenient and manageable subtasks. Each of the modules can be written and tested in isolation. Only when the grammarian is satisfied will he proceed with the description of another (set of) structure(s). The modular structure is also found to be very practical in case one wants to compare alternative descriptions. While the rest of the grammar is left unchanged, one module may be exchanged for another. In a similar fashion, of course, minor changes and/or additions can be put into effect, without having to compile the entire grammar all over again. From the point of view of efficiency the modular structure thus also proves to be rather attractive since it allows for the partial re-compilation of a grammar. Consequently, efficiency is increased and the cost involved -- in terms of computer time -- reduced.

The two levels of the extended affix grammar were each given a distinct role. Thus the context-free level is used to describe the constituent structure of strings in terms of functions and categories. The affix-level is occupied by two types of attribute: syntactico-semantic affixes and what may be referred to as "steering" affixes. Syntactico-semantic affixes typically serve to supply additional information about (the elements of) the structure and make it possible to establish whether certain relations hold between constituents (e.g. concord). Steering affixes, on the other hand, merely serve to impose restrictions on the application of particular rules. Thus, for example, a steering affix may be used to record what part of the input string was recognized, or what branching was postulated for a particular node. Steering affixes, although linguistically less relevant, can effect a reduction in the amount of ambiguity that would result from an unrestricted application of the rules.

4 The Grammar: A formalization of descriptive rules

4.1 Introductory

While a discussion of the design of the grammar was presented in the previous chapter, together with a general outline of the structure of the grammar, the present chapter deals more specifically with the description of coordination and gapping, and the noun phrase. The motivation for including a discussion of the phenomena of coordination and gapping can be found in the fact that these account for the multi-layered structure that was introduced throughout the grammar. A discussion of parts of the formalized description of the noun phrase forms an example of what such a description actually looks like.[1] Moreover, it enables us to illustrate how the description of coordination may be integrated with the description of a phrasal category like the noun phrase.

4.2 Coordination and gapping

In this section a more detailed account is given of the general structure of the grammar at sentence and clause level. More specifically we include a discussion of some of the working-principles involved in the handling of coordination and gapping.[2] For the sake of discussion we restrict ourselves here to an illustration of these principles as they apply to the structure that is assigned to the (regular) declarative sentence. Our grammar, however, accounts for both instances of backward conjunction reduction and forward conjunction reduction occurring not only with regular declarative sentences, i.e. sentences with an unmarked wordorder, but also with various other types of sentence such as extraposed, existential and cleft sentences whether declarative or interrogative.[3]

---

[1] We do not give a full description of the noun phrase because this would involve a discussion of the grammar as a whole.

[2] Part of this section was originally published in Aarts and Meijs (eds.) (1986): 177-202.

[3] See also Oostdijk (to appear).

As was observed in section 3.5 we opted for a description in terms of immediate constituents (ICs), incorporating both functions and categories. An analysis will thus yield a multi-layered constituent structure. We proceeded by writing various subgrammars, taking, in general, the traditional phrases as a starting-point. At a later stage these subgrammars were combined so as to form one grammar with a clear modular structure. For the description of each of the phrases and also at sentence level, we introduced a multi-layered structure rather than what could be referred to as a flat one, as found in for instance Quirk et al. (1985, 1972), and Aarts and Aarts (1982). Although the structures found in these handbooks are 'multi-layered' in the sense that they are IC-based and represent the rank hierarchy, we call them flat, because they also reflect the linear order of constituents. A major drawback of such an approach is that coordination can only be accounted for in terms of phrases or clauses or their immediate constituents. This entails that types of coordination which do not involve 'full' phrases or clauses can only be accounted for by postulating various kinds of ellipsis. In writing a formal grammar it appears extremely complex to keep track of all information concerning the presence or absence of elements. Our principal argument then for using a multi-layered structure in which function slots are structurally ordered with respect to one another is the fact that in doing so we are able to account for any type of coordination without having to postulate elliptic structures in almost every instance. In designing a multi-layered structure we have found that coordination can be optimally described by having, preferably, binary branching nodes and possibly single branching nodes, but never multiple branching beyond binary, that is, when taking only obligatory constituents into consideration. This can be illustrated by considering the sentence structure associated with declarative sentences.

Taking the clause types described by Quirk et al. (1985: 719ff; 1972: 342ff)[4] as a starting-point we set ourselves the task of designing a structure that will accommodate each of the following types:

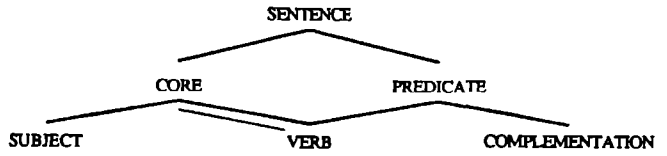
[4] Note that type (6) has been included as a basic type in addition to the types distinguished by Quirk et al. (1985: 721; 1972: 343-344). Unlike in Quirk et al., the two types that include an obligatory adverbial constituent (SU-VB-A and SU-VB-OD-A) have as such been left out; instead, structures of this kind as presented by Quirk et al. are accounted for in terms of the types SU-VB-CS and SU-VB-OD-CO respectively.

1. SU - VB (intransitive)
2. SU - VB (intensive) - CS
3. SU - VB (monotransitive) - OD
4. SU - VB (ditransitive) - OI - OD
5. SU - VB (complex transitive) - OD - CO
6. SU - VB (complex ditransitive) - OI - OD - CO

In addition, any extensions of these six basic types by the insertion of optional adverbials must be accounted for.

In Figure 1 a graphic representation of the structure of the grammar on this point is given, not to be confused with a derivation tree.

Figure 1



The figure might suggest that the verb can be dominated by two mother nodes at the same time. This is not the case, however. It will be seen that the domination by the CORE node is indicated by a double line, that by the PREDICATE node by a single line. The difference indicates that VERB is dominated *either* by CORE *or* by PREDICATE. The domination by CORE (the double line) is only activated under condition of coordination, while the single branch reflects the domination that holds where no coordination is present. In this way the node immediately dominating the subject and the verb allows for the coordination found in (1)-(3) to be analyzed as the coordination of cores. The double line indicates that, in case there is no coordination of cores, the core will only dominate the subject, while the verb is dominated by the predicate.

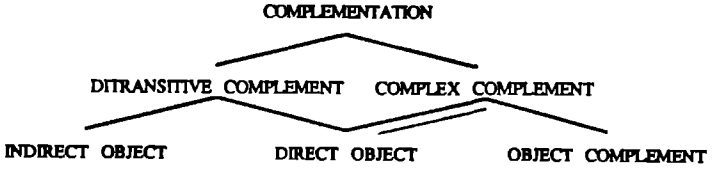
(1) Sir John proposed and Sir Humphrey seconded the motion.
(2) He pushed and his sister pulled the boat ashore.
(3) Paul offered but Matthew actually handed me his coat.

The internal structure of the complementation varies according to the verb. In the case of an intensive or monotransitive verb, for example, the complementation node is a single branching node (disregarding any optional adverbials for the time being). Complex transitive and ditransitive verbs as well as complex ditransitive verbs have as sister nodes binary branching complementation nodes. In the case of complex ditransitive verbs, i.e. verbs requiring an indirect and a direct object as well as an object complement, we make use of the branching for ditransitives and complex transitives, so that a node 'ditransitive complement' dominates the indirect object and the direct object, and a node 'complex complement' dominates the direct object and the object complement. With intransitive verbs there is no complementation. The different structures for the complementation are represented in the tree diagrams below (Figures 2-4).

Figure 2: Intensive and monotransitive complementation

Figure 3: Ditransitive and complex transitive complementation

Figure 4: Complex ditransitive complementation



The structure associated with complex ditransitive complementation shows a direct object that can be dominated by a ditransitive complement and a complex complement. The double line indicates the fact that if the complementation has a minimal realization, i.e. just an indirect object followed by a direct object and an object complement, the direct object will be considered to be part of the ditransitive complement only. This decision is motivated by the fact that the object complement in such structures is almost always optional, and then behaves as an optional constituent, while the verb can be looked upon as ditransitive. Consider the following examples:[5]

(4) He caught me the rabbit alive.
(5) She returned me my book covered with stains.

Dominance of the direct object by the complex complement rather than the ditransitive complement will be assumed in cases where an indirect object is followed by a coordination of a maximum complex complement, i.e. a direct object followed by an object complement, as in examples (6) and (7) below.

(6) He poured us the wine ice-cold and the beer lukewarm.
(7) The postman delivered me the books undamaged but the magazines torn.

A co-occurrent dominance of the direct object by the ditransitive complement can be found in sentences like (8) and (9), where we have coordination of ditransitive complements followed by a complex complement or a coordination of complex complements.[6]

(8) They sold us a car and my brother a motor-cycle without spare tyres.
(9) She poured him a wodka and his friend a gin, the one on the rocks and the other straight.

---

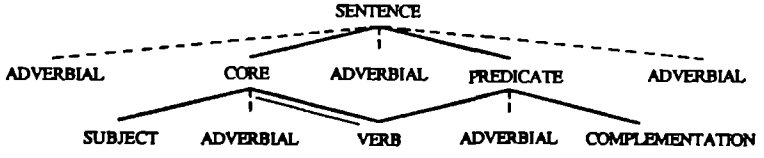
[5] Note that without further context the examples must be taken to be ambiguous. The intended interpretation here, however, requires the assignment of the structure SU-VB-OI-OD-CO.

[6] It should be clear that while examples (8) and (9) might be interpreted in a way so that *my brother* (8) and *his friend* (9) are taken to function as subjects, this reading is not intended here.

Reconsidering the various structures assigned to the complementation it appears worthwhile to combine these into one structure that will accommodate each of the types. Therefore in line with the structure of the complementation assumed for complex ditransitive complementation, we introduce an intermediate level between the complementation and the objects and the object complement, also in the case of ditransitive or complex transitive complementation, so that they will be dominated by a ditransitive and a complex complement at all times. The subject complement is assumed to be immediately dominated by the complementation node.

Allowing for any optional adverbials to occur in various positions we can now extend our structure for the sentence as in Figure 5. This, again, is a graphical representation of the structure of the grammar, not a derivation tree. Whether the adverbials preceding and following the verb will actually occur as valid optional constituents in a structure will depend on the presence or absence of the branches core-verb and predicate-verb. These in turn are conditioned by the coordination of cores. If the core is not coordinated, the verb will only be dominated by the predicate and the core node will only dominate the subject (as represented in Figure 5a). In other words, if the core is not coordinated, any adverbial preceding the verb will be dominated by the sentence node and no ambiguity will arise. However, if the core is coordinated, the core node will dominate both the subject and the verb and also, possibly, any adverbials immediately preceding the verb (consider Figure 5b). With the coordination of cores the predicate-verb branch no longer occurs and the predicate node only dominates the complementation.

Figure 5



A similar situation is found in the case of the complementation structures. With intensive and monotransitive verbs we assume the

Figure 5a

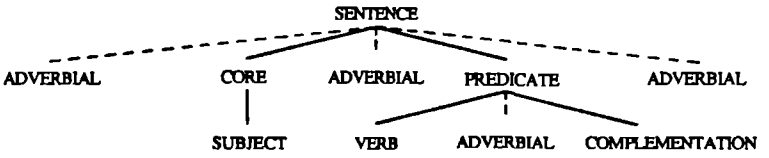
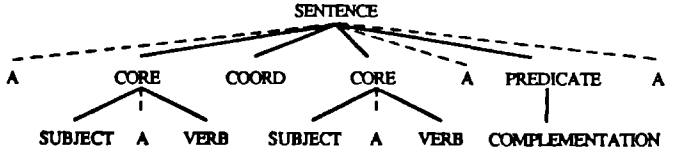


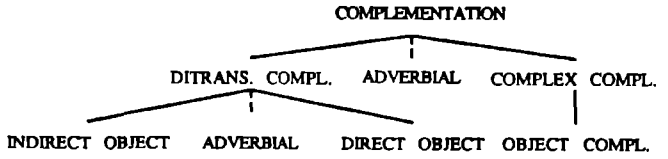
Figure 5b



complementation node to dominate a single obligatory constituent. In order to avoid any ambiguity in the analysis of adverbials in such structures we do not allow for any adverbials to occur under dominance of the complementation node. With ditransitives, complex transitives and complex ditransitives the possible occurrence of adverbials dominated by the nodes of ditransitive complement, complex complement and complementation respectively is conditioned by the branches chosen. Thus in the case of a ditransitive verb the complementation node dominates only the ditransitive complement, in the case of a complex transitive verb just the complex complement, and only in the case of a complex ditransitive verb both the ditransitive complement and the complex complement as well as, possibly, any adverbials.

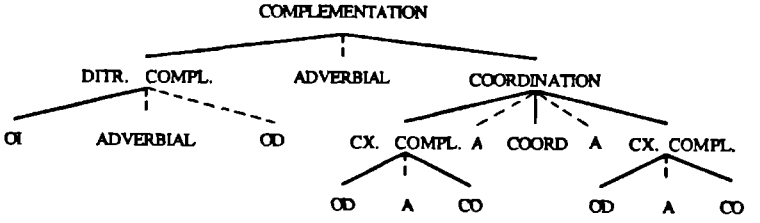
The ditransitive complement node and the complex complement node may each dominate any adverbials between the obligatory constituents. Not always so, however, in the case of a complex ditransitive verb. Here again the grammar allows for conditional branches: in the case of a complex transitive, under the condition that there is no coordination of the complex complement, the direct object will be dominated by the ditransitive complement, in which case any adverbial(s) immediately preceding the direct object will be dominated by the ditransitive complement (Figure 6).

Figure 6



With the coordination of the complex complement the direct object is always dominated by the complex complement node although it also still occurs under the ditransitive complement node as an optional constituent (Figure 7).

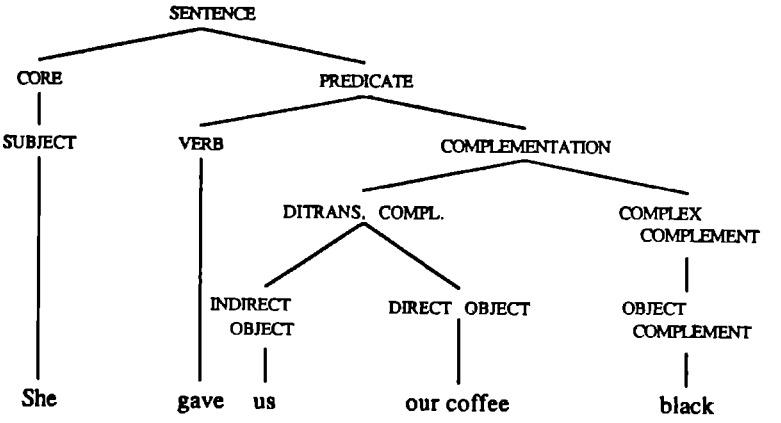
Figure 7



The adverbial option following the direct object is a true option, whereas the adverbial option immediately following the indirect object is conditioned by the ditransitive complement dominance of the direct object. Note that at all levels in the grammar but the highest, adverbials are found in a binary branching environment, i.e. we allow for adverbials to occur in positions where both the adjacent sister node(s) to the left and the adjacent sister node(s) to the right are dominated by one and the same mother node. This entails that adverbial options may also occur under the influence of coordination: by coordinating two or more constituents new binary branching nodes are created.

The structure for the declarative sentence as outlined above allows for coordination in a great many positions. In fact, any constituent can be coordinated with a similar constituent, i.e. a constituent containing the same functions in the same order. For example, given the sentence 'She gave us our coffee black' any coordinated variants of this sentence can be accounted for by extending the tree structure given below (Figure 8).

Figure 8



Subjects, predicates, cores, verbs, complementations, objects and complements, can all be coordinated. Possible extensions of the sentence above include, for example, the following:

(10) She and her sister gave us our coffee black.
(11) She gave us and her father our coffee black.
(12) She gave us our coffee black and her mother hers white.
(13) She gave us our coffee black and our tea white.

Sentence (10) is an example of coordination of the subject; in (11) the indirect object is coordinated; in (12) we have coordination of the complementation; and in (13) the complex complement is coordinated. The derivation trees for these examples are given in Figures 9-12.

Figure 9

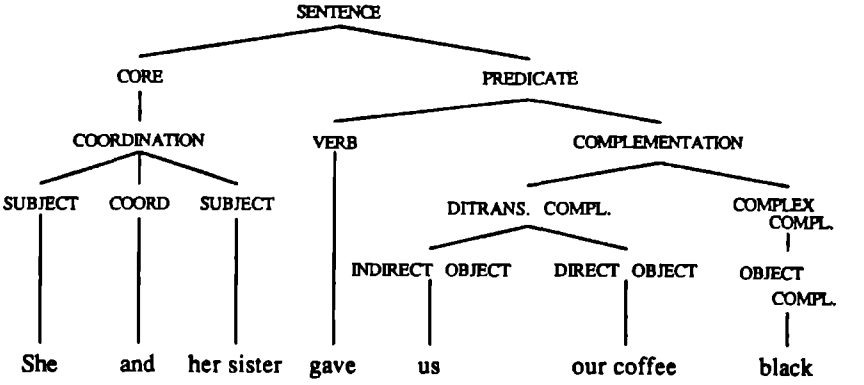


Figure 10

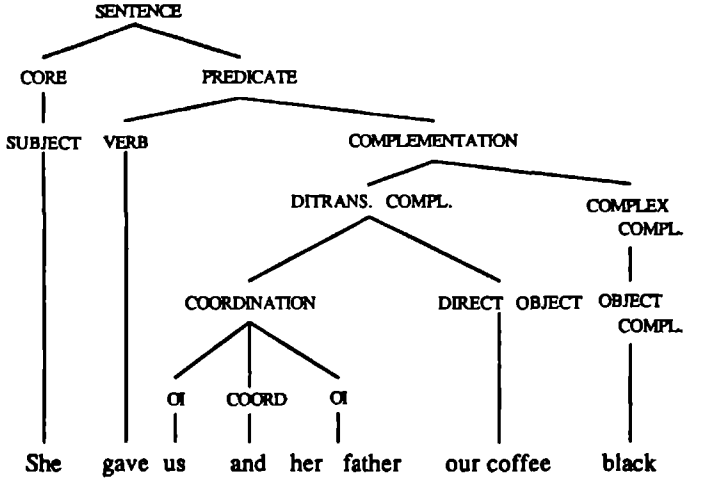


Figure 11

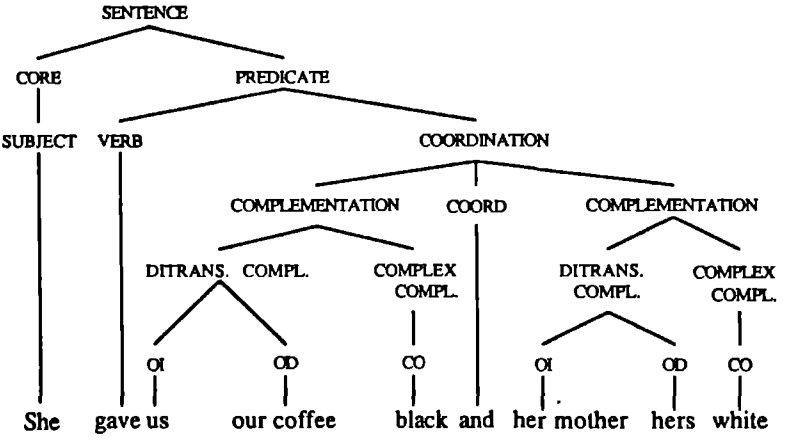


Figure 12

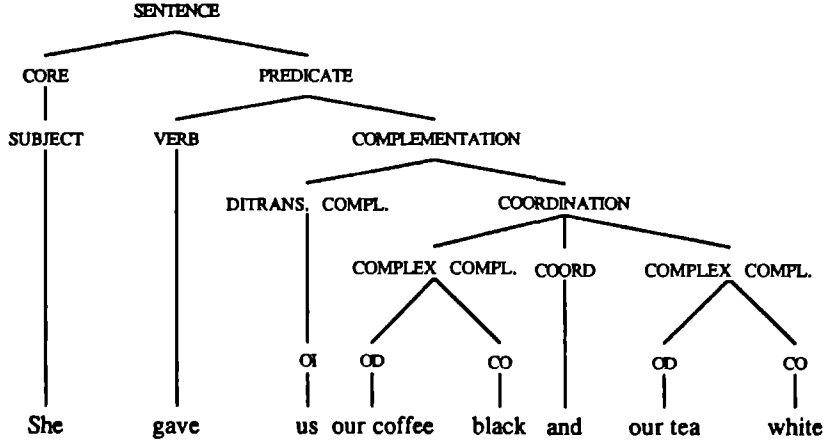
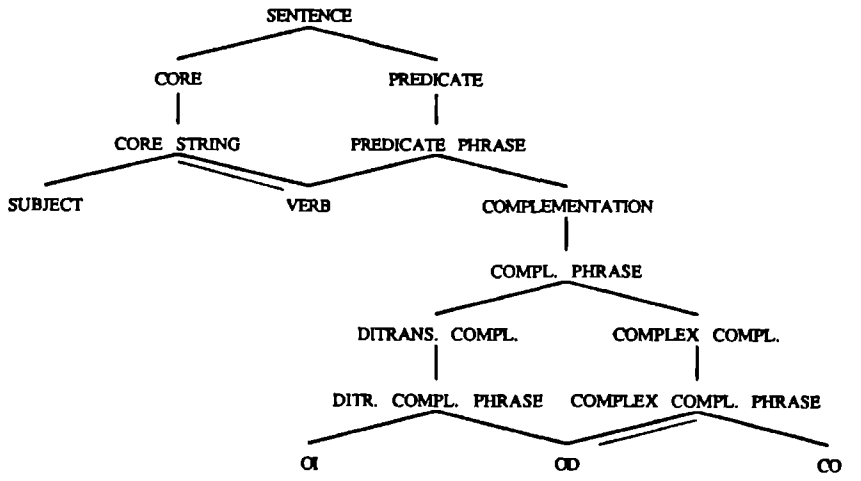


Figure 13



At this point it should be observed that in designing a multi-layered structure for the declarative sentence we have been (primarily) concerned with the placement of sentence functions. No attention has been given to the desired alternation between functional and categorial labels. The introduction of the required categorial labels yields a further structuring of the description of the declarative sentence and can be represented as in Figure 13.

In the examples above we have assumed that instances of coordination always involve the coordination of functional constituents. It seems, however, arbitrary to look upon (10) and (11) as the coordination of subjects and indirect objects respectively: they might as well be analyzed in terms of the coordination of noun phrases. It could even be argued that there are certain coordinations of noun phrases that cannot be accounted for in terms of the coordination of a particular function. This is the case, for instance, with noun phrases sharing a postmodifier as in examples (14) and (15).

(14) *The young chief of staff and his senior colleague who attended the conference* returned to Washington to report to the President.
(15) *The boy and his sister whose parents had been killed in the accident* had to go to the orphanage.

If the relative clauses are taken to postmodify both noun phrases in the sentences it appears impossible to account for the coordination in terms of the coordination of subjects. This leads us to conclude that coordination should rather be accounted for in terms of coordination of (identical) categorial constituents. Indeed, the coordination of identical categorial constituents presents no problems. We may, however, also come across coordinations involving different categories (see 16-18):

(16) He had forgotten *where and at what time* he was expected to meet him.
(17) *Bored and without anything else to do* he decided to go for a walk.
(18) A man, *nearly starving and without any energy left*, came to the village after having wandered through the wilderness for days.

In (16)-(18) different categorial constituents are coordinated. This seems to suggest that category information by itself is not sufficient to base a description of coordination upon. Describing coordination solely in terms of functions we consider most unattractive, since, in addition to the argument of noun phrase coordination above, it would entail an overgeneralization: such a description would suggest that the choice with respect to the realization of functional constituents (by categorial constituents) is free. This is, of course, not the case, as can easily be demonstrated by means of the following examples:

(19) * His brother is *a good man and in the United States.*
(20) * *Going away like that and that he had not said anything about it* had offended her.

These and other instances clearly show that there must be restrictions on the choice of categorial constituents for the coordination of certain functions.

From what we have observed above it may be inferred that neither a description of coordination based solely on categories nor one based exclusively on functions is satisfactory; rather it is suggested that any description of coordination should be based on both functions and categories. Since the coordination of identical categories (on the condition that they are dominated by one and the same function node) is always possible and the coordination of different categories only in some cases, the description of coordination should basically be category-based and supplemented by information concerning the function of the node immediately dominating the coordinated (categorial) constituents.

We therefore take it as a working-principle that coordination involves (minimally) two categorial constituents dominated by one and the same function node. Coordination occurs throughout the grammar and is what we might call a process in the grammar. Unlike the modules that describe phrasal and similar categories and thus form more or less self-contained entities, processes are not restricted to one particular module, but are relevant to several or even all modules. Processes can be looked upon as 'subroutines' in that it is possible to formulate rule-schemata which, when called upon, will operate according to the rule-generating rule principle.

The description of coordination is formulated in generalized rules requiring as input (minimally) two categorial constituents. An example of the description of syndetic coordination might look as follows:[7]

(rewrite rule 1)

SYNDETIC COORDINATION (function + poss info):
CONJOIN (category 1, function + poss info 1),
CONSTITUENT SEQUENCE (category 2, function + poss info 2),
identical or different categories (category 1, category 2, function),
any additional info (function, poss info 1, poss info 2, poss info).

[7] Here we present the rules that were formulated for 'higher level' coordination, i.e. the coordination found at a level higher than the traditional phrase. For the description of coordination found within the NP, VP, AJP, AVP and PP a separate (though similar) set of rules was introduced since at this level adverbial and connective elements do not occur. Apart from avoiding the overhead that would be introduced by having a set of fully generalized rules, i.e. one set of rules to account for all coordination irrespective of what level it is found at, the distinction between the two 'levels' of coordination is further motivated by the fact that at the 'higher' level correlative coordination may occur freely, while at the 'lower' level its occurrence is restricted. See also section 4.3.

(rewrite rule 2)

```
CONSTITUENT SEQUENCE (category, function + poss info):
   ADVERBIAL OPTION,
     COORDINATOR(type),
       CONNECTOR OPTION(type),
         ADVERBIAL OPTION,
           CONJOIN (category 1, function + poss info 1),
             CONNECTOR OPTION (type),
               ADVERBIAL OPTION,
                 MORE CONSTITUENTS (category 2, function +
                                                    poss info 2),
                   identical or different categories (category 1, category 2,
                                                        function),
                     any additional info (function, poss info 1,
                                                 poss info 2, poss info).
```

(rewrite rule 3)

```
MORE CONSTITUENTS (category, function + poss info):
     ;
   CONSTITUENT SEQUENCE (category, function + poss info).
```

With each constituent a category, a function and possibly additional information (concerning, for example, concord or complementation) is associated. In formulating an implicitly equal function, we require the coordination of two constituents to be dominated by one and the same function node. The first predicate (identical or different categories ...) allows for identical categories to be coordinated in case they are dominated by the same function (rewrite rule 4), but also for different categories to be coordinated, given the dominance by a specific function (rewrite rule 5).

(rewrite rule 4)

```
identical or different categories (category, category, function): .
```

(rewrite rule 5)

```
identical or different categories ("NP", "PP", "A"): .
```

In rule (5) only one example is given. Rules similar to this rule need to be added for every instance where coordination of different categories is possible.

Whereas the first predicate clearly imposes restrictions on the application of the rules for coordination, the second predicate (any additional info ...) may function in two ways: not only can it impose restrictions on the application of the rules, it is also possible to control the effect of coordination on matters like concord.

An example of imposing restrictions on the application of the coordination rules can be found in rules (6) and (7):

(rewrite rule 6)

any additional info (function, complementation, complementation,
complementation): .

(rewrite rule 7)

any additional info ("PREDICATE", complementation 1, complementation 2,
"DIFFERENT COMPLEMENTATION"):
not equal (complementation 1, complementation 2).

Rewrite rule (6) requires the complementation in a coordination, irrespective of the function it is associated with, to be identical. Rewrite rule (7) concerns an exception to rule (6) and allows for constituents with different complementations to be coordinated if the coordination is dominated by the predicate.

The effects of coordination on concord can be expressed in the same predicate rule. Thus for 'additive coordination', i.e. coordination by means of the coordinator *and* we can formulate the following rules:

(rewrite rule 8)

any additional info (function, number 1 + person 1, number 2 + person 2,
"PLU" + person):
additive person (person 1, person 2, person).

(rewrite rule 9)

```
additive person (person, person, person): .
```

(rewrite rule 10)

```
additive person ("1ST", 2nd or 3rd, "1ST"): .
```

(rewrite rule 11)

```
additive person (2nd or 3rd, "1ST", "1ST"): .
```

(rewrite rule 12)

```
additive person ("2ND", "3RD", "2ND"): .
```

(rewrite rule 13)

```
additive person ("3RD", "2ND, "2ND"): .
```

Any additive coordination results in a plural ("PLU") number, while the effect on person is specified in rewrite rules (9)-(13): if the values for person of each of the constituents differ, the resultative, additive person will fall in with the 'lowest' value found for person in the members of the coordination; if the values for person correspond, the additive person will have the same value.

The description of coordination as a process and the formulation of rule schemata have a maximum effect when we can work with generalized non-terminals, such as CONJOIN. As a side-effect we must introduce what we shall call 'linking rules' in the various modules, i.e. rules that will call upon the rules for coordination. We therefore formulate rules like (14) and (15):

(rewrite rule 14)

```
VERB (concord, complementation):
    VERB PHRASE (concord, complementation);
    COORDINATION ("VB" + concord + complementation).
```

(rewrite rule 15)

CONJOIN ("VP", "VB" + concord + complementation):
    VERB PHRASE (concord, complementation).

so that a function verb ("VB") will be realized by either a single verb phrase or, in case the coordination rules are called upon, a coordination of verb phrases.

As a general strategy in the formulation of linking rules we assume that a (categorial) constituent, except for the lexical categories, must dominate at least two obligatory (functional) constituents in order to be able to occur in a coordination. This means that if a higher level categorial constituent dominates a single function, which in turn dominates a single category, there will be no ambiguity as to what level the coordination must be associated with. For example, a coordination as found in (21) will be looked upon as the coordination of nouns rather than of NPs.

(21) John and Peter came to her birthday.

At this point we may observe that the introduction of a multi-layered structure and the description of coordination as a process have made it possible to handle many instances of coordination in an efficient manner on the basis of elements that are actually present, without having to postulate elliptic structures in almost every instance. Unfortunately, however, not all instances of coordination can be accounted for in this way.

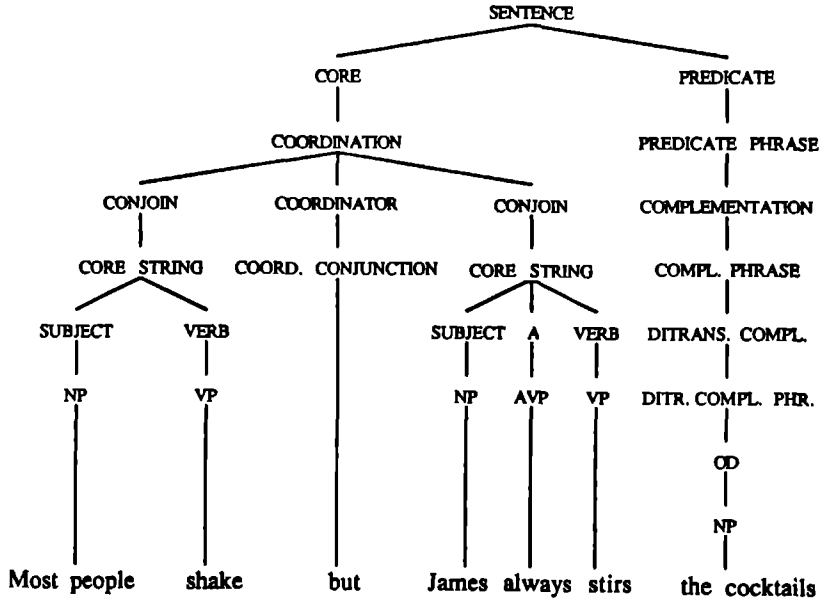
In the approach described above, instances involving the coordination of full conjoins are never problematic. Since we make use of a multi-layered structure, many instances involving either what is usually referred to as backward conjunction reduction (i.e. the coordination of a reduced first conjoin and a full second conjoin) or forward conjunction reduction (i.e. coordination of a full first conjoin and a reduced second conjoin) can be handled satisfactorily as coordinations of 'full' conjoins, i.e. full conjoins in terms of our grammar. Thus instances of coordination as found in (22)-(25) can be dealt with in this manner.[8]

[8] Note that without further context (25) is ambiguous. The intended reading here assumes *Jim* to be an indirect object (and not a subject).

(22) Most people shake but James always stirs the cocktails.
(23) He read the letter and laughed.
(24) John attended the course regularly and passed the exam.
(25) She gave Jack two letters and Jim the telegram stained with coffee.

The derivation trees for these structures can be found in Figures 14-17.[9]

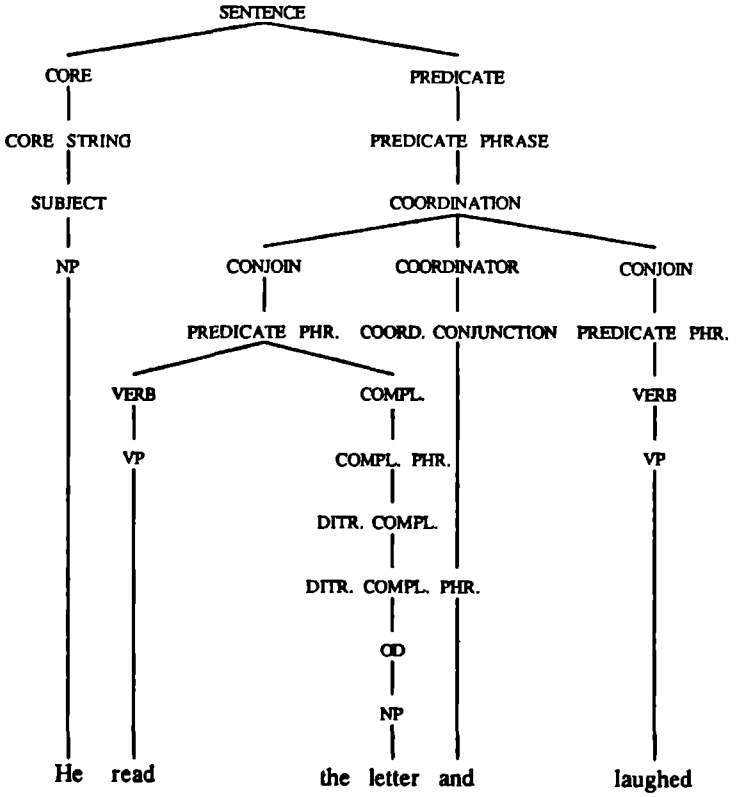
Figure 14



Some instances showing backward conjunction reduction, however, remain problematic, even when we make use of a multi-layered structure, since the coordination found in such instances cannot be accounted for in terms of the coordination of full conjoins at whatever level.

[9] In the derivation trees below we have 'filtered out' the intermediate labelling of constituents in coordinations, such as CONSTITUENT SEQUENCE and MORE CONSTITUENTS.

Figure 15

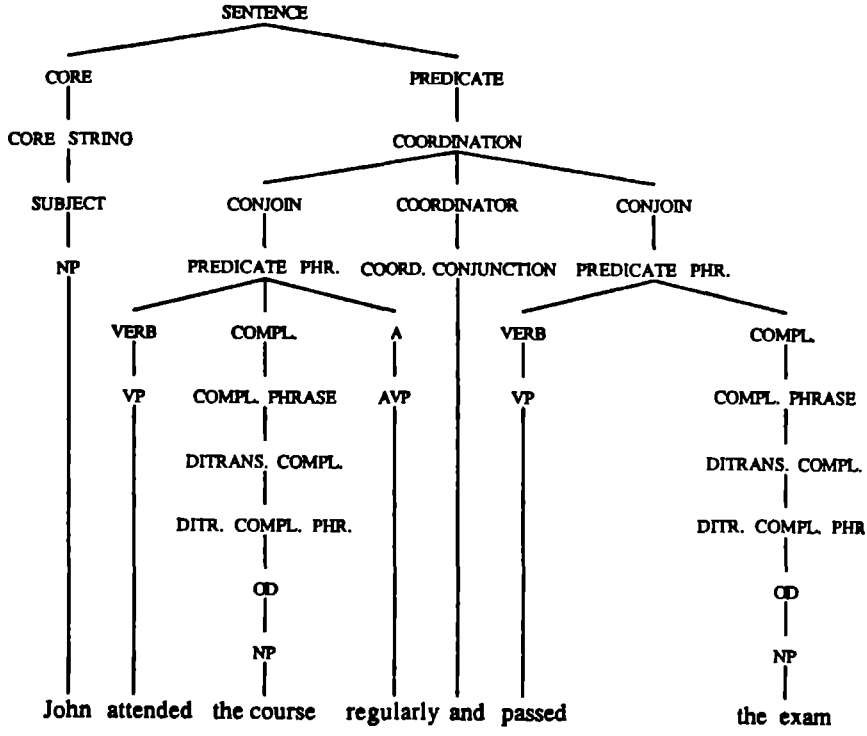


Consider the following examples (Bresnan, 1974: 618):

(26) I can tell you when /, but I can't tell you why he left me.
(27) I've been wondering whether /, but / wouldn't positively want to state that, your theory is correct.

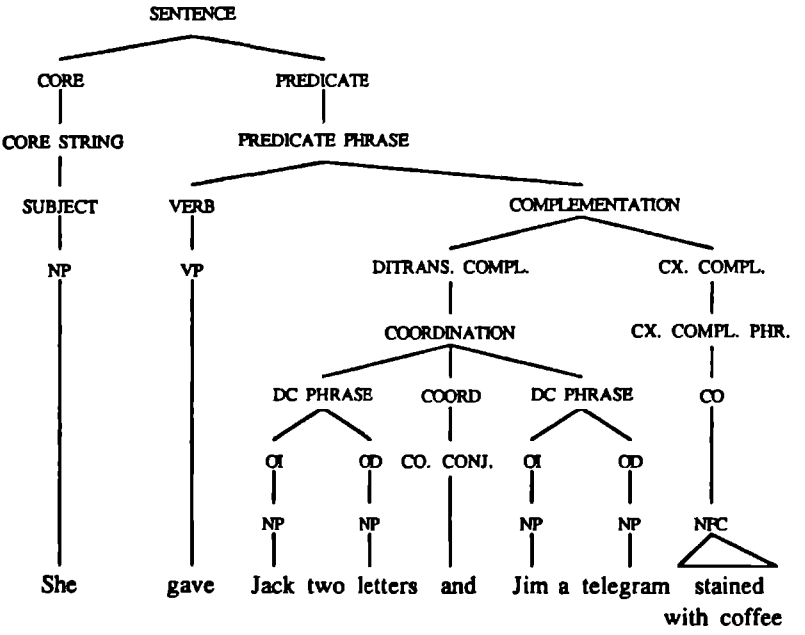
Looking at the phrase marker for each of these sentences it is clear where the problem lies with respect to the description of this kind of coordination (see Figures 18 and 19).

Figure 16



In example (26) the coordination occurs at sentence level, while in (27) the predicate is coordinated. In both cases we are confronted with a situation where the level of coordination is determined by the second, full conjoin, while at the same time the first conjoin is reduced at a point which we cannot account for in the present structure. The description of such instances requires a set of additional rules specific for coordinations like these. Since we do not expect them to be very frequent we have, so far, refrained from incorporating these in the grammar.

Figure 17



Other instances of coordination that have not yet been accounted for in the approach described above include the following:

(28) This man is a doctor and his wife a teacher.
(29) He gave him the coffee and the woman the cake.

Neither of the structures found in (28) and (29) can be accounted for in terms of the structure assigned to the declarative sentence, that is, if -- as we intend to do here -- the clause pattern found in these sentences is taken to be SU-VB-CS-SU-CS and SU-VB-OI-OD-SU-OD respectively.[10] In (28) it is the absence of the verb in the second conjoin which blocks a possible analysis as the coordination of sentences, whereas at the same time no other analyses are possible, since there is no possibili-

[10] Note that the interpretation of (29) as SU-VB-OI-OD-OI-OD is not problematic since this can be accounted for in terms of the coordination of ditransitive complements.

Figure 18

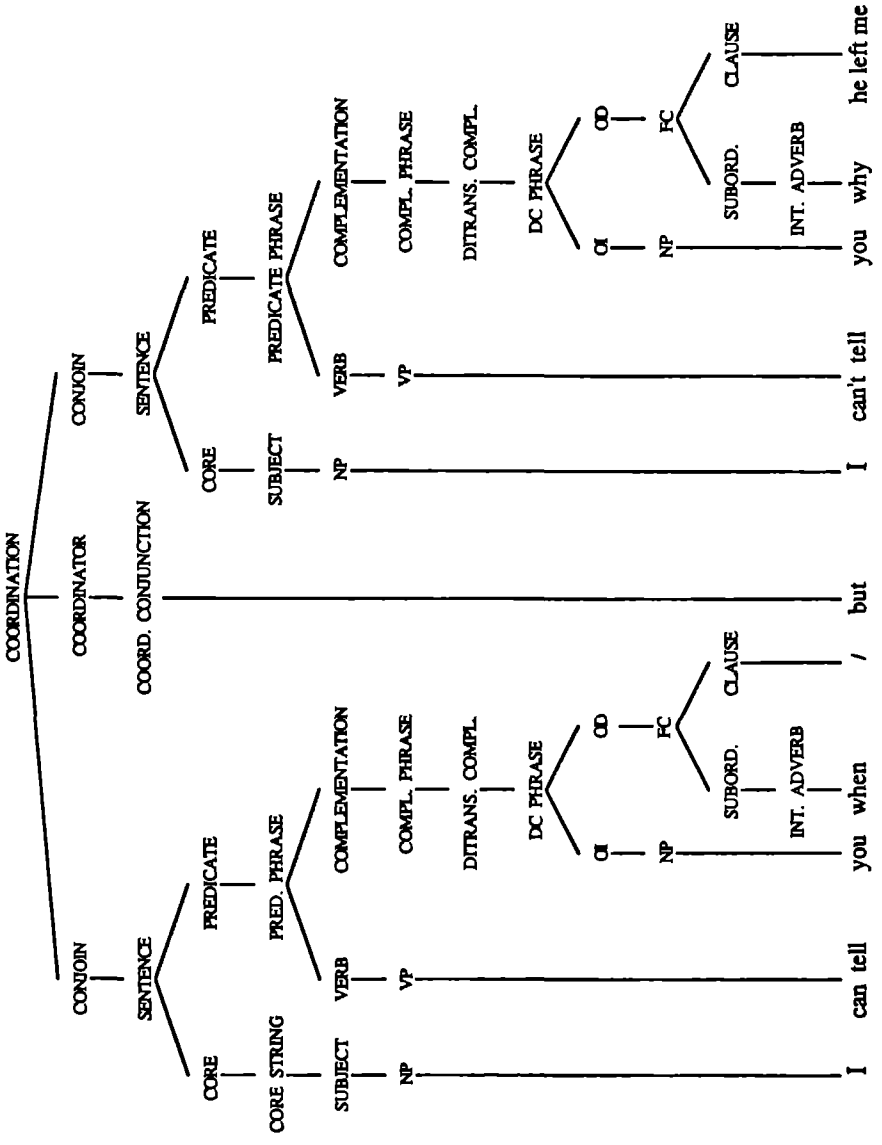
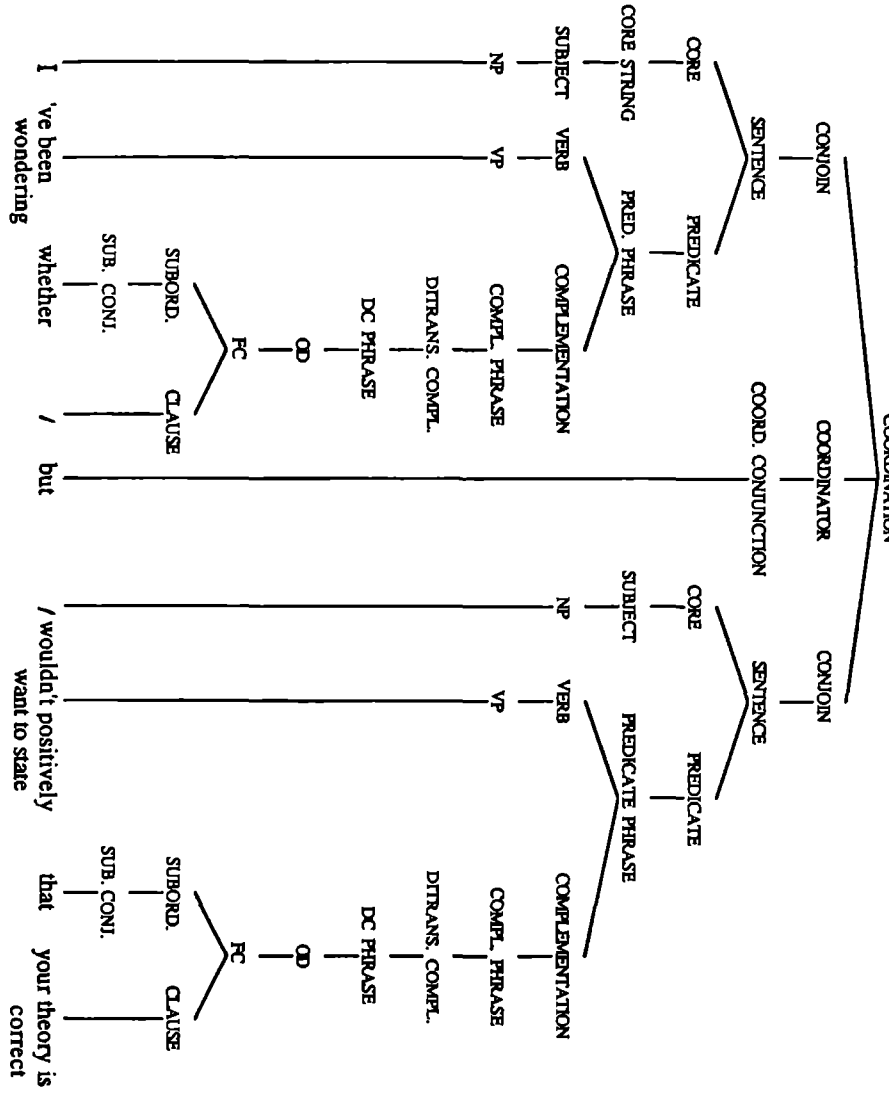
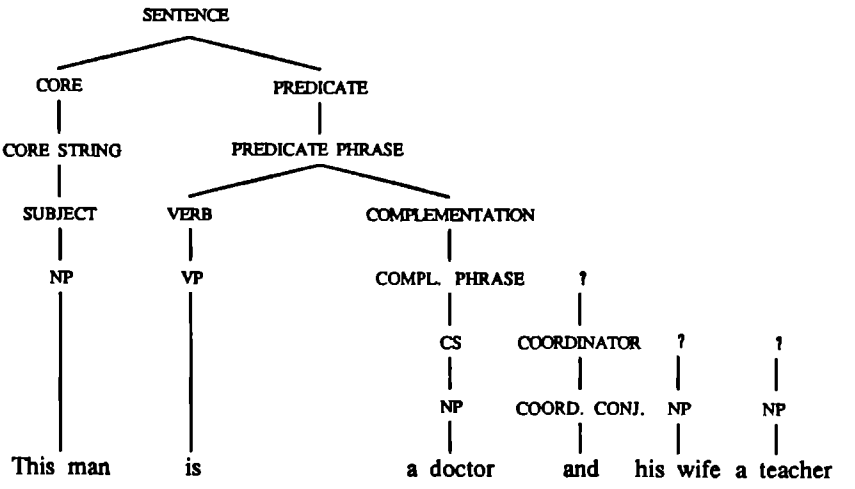

COORDINATION
CONJOIN
SENTENCE
CORE
CORE STRING
SUBJECT
NP
I
PREDICATE
PRED. PHRASE
VERB
can tell
VP
COMPLEMENTATION
COMPL. PHRASE
DITRANS. COMPL.
DC PHRASE
OI
NP
you
OD
FC
SUBORD. CLAUSE
INT. ADVERB
when
COORDINATOR
COORD. CONJUNCTION
but
CONJOIN
SENTENCE
CORE
SUBJECT
NP
I
PREDICATE
PREDICATE PHRASE
VERB
can't tell
VP
COMPLEMENTATION
COMPL. PHRASE
DITRANS. COMPL.
DC PHRASE
OI
NP
you
OD
FC
SUBORD. CLAUSE
INT. ADVERB
why
he left me

Figure 19


COORDINATION
CONJOIN
SENTENCE
CORE
CORE STRING
SUBJECT
NP
I
PREDICATE
PRED. PHRASE
VERB
VP
've been wondering
COMPLEMENTATION
COMPL. PHRASE
DITRANS. COMPL.
DC PHRASE
OD
FC
SUBORD. CLAUSE
SUB. CONJ.
whether
COORDINATOR
/
COORD. CONJUNCTION
but
CONJOIN
SENTENCE
CORE
SUBJECT
NP
/ wouldn't
PREDICATE
PREDICATE PHRASE
VERB
VP
positively want to state
COMPLEMENTATION
COMPL. PHRASE
DITRANS. COMPL.
DC PHRASE
OD
FC
SUBORD. CLAUSE
SUB. CONJ.
that
your theory is correct

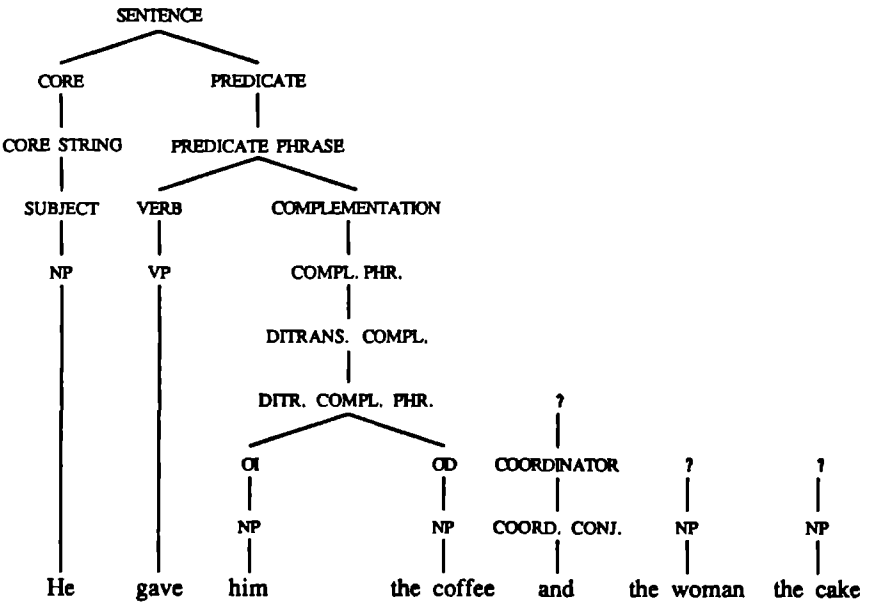
ty for placing the coordination at any other level allowing for an analysis in terms of the coordination of full conjoins (cf. Figure 20). A similar situation is found in (29) where not only the verb is missing but also the indirect object (cf. Figure 21).

Figure 20



Contrary to the instances of backward conjunction reduction we discussed above instances like the ones in (28) and (29) are rather frequent. They are instances of gapping, i.e. they are typically instances of coordination at sentence level involving (minimally) two conjoins, the first of which has a full realization while in the second the verb and possibly other elements are found to be missing. Here the condition holds that no elements may be assumed missing unless they are found in the first conjoin. Whereas the coordination of full conjoins can be looked upon as a general process, the coordination found in cases of gapping appears to be restricted to sentence level. Given this and the fact that obviously a description of gapping cannot be achieved without having to allow for elements to be missing, we suggest that a special provision should be created in order to be able to account for instances of gapping.

Figure 21



Although the gapping we want to describe is restricted to sentence level (i.e. the coordination is found on sentence level) its description affects the entire grammar, since record must be kept of what elements are missing at what points, whereas before we only had to allow for full constituents with occasional exceptions in the case of conditioned optionality. Considering once more the basic types we distinguished above (cf. p. 83), we set ourselves the task of describing the structures listed in Table 1 on page 107 as instances of gapping.

The distribution of the basic sentence functions over the multi-layered structure forces us to pass along the information concerning the presence or absence of a certain constituent over various levels. This is done by means of an affix flow holding this type of information. What constituents may be missing at what point is stated in various predicate rules. Possible ellipsis is considered per level. Below some rules from the grammar are given by way of illustration.

	coordinate structure	
complement. type	1st conjoin (full)	2nd conjoin (red.)
intransitive	SU-VB	SU
intensive	SU-VB-CS	SU
	SU-VB-CS	SU-CS
monotransitive	SU-VB-OD	SU
	SU-VB-OD	SU-OD
ditransitive	SU-VB-OI-OD	SU
	SU-VB-OI-OD	OI
	SU-VB-OI-OD	SU-OI
	SU-VB-OI-OD	SU-OD
	SU-VB-OI-OD	SU-OI-OD
complex transitive	SU-VB-OD-CO	SU
	SU-VB-OD-CO	OD
	SU-VB-OD-CO	SU-OD
	SU-VB-OD-CO	SU-CO
	SU-VB-OD-CO	SU-OD-CO
complex ditransitive	SU-VB-OI-OD-CO	SU
	SU-VB-OI-OD-CO	OI
	SU-VB-OI-OD-CO	OD
	SU-VB-OI-OD-CO	SU-OI
	SU-VB-OI-OD-CO	SU-OD
	SU-VB-OI-OD-CO	SU-CO
	SU-VB-OI-OD-CO	SU-OI-OD
	SU-VB-OI-OD-CO	SU-OD-CO
	SU-VB-OI-OD-CO	SU-OI-CO
	SU-VB-OI-OD-CO	SU-OI-OD-CO

arting from the beginning of the grammar we find a number of rules scribing the possible realization of the function utterance:

(rewrite rule 16)

```
UTTERANCE:
    SENTENCE (complementation, "FULL");
    COORDINATION OR GAPPING ("UTT" + complementation + realization).
```

(rewrite rule 17)

```
COORDINATION OR GAPPING ("UTT" + complementation + realization).
    CONJOIN ("S", "UTT" + complementation 1 + realization 1),
      COORDINATOR (type),
        CONNECTOR OPTION (type),
          CONJOIN("S", "UTT" + complementation 2 + realization 2),
            CONNECTOR OPTION (type),
              gapping or full coordination (complementation 1,
                                    complementation 2, realization 2).
```

(rewrite rule 18)

```
CONJOIN("S", "UTT" + complementation + realization):
    SENTENCE (complementation, realization).
```

(rewrite rule 19a)

```
gapping or full coordination ("INTRANSITIVE", "INTRANSITIVE",
                                                    "FULL"): .
```

(rewrite rule 19b)

```
gapping or full coordination ("INTRANSITIVE", "INTRANSITIVE",
                                                "MISSING PREDICATE"): .
```

(rewrite rule 19c)

```
gapping or full coordination (complementation, complementation, realization):
    not equal (complementation, "INTRANSITIVE").
```

(rewrite rule 19d)

```
gapping or full coordination (complementation 1, complementation 2, "FULL"):
    not equal (complementation 1, complementation 2).
```

By means of the first alternative of rule (16) those instances are described in which there is no coordination at sentence level and an utterance is realized by a single 'full' sentence. The application of the second alternative of rule (16) in combination with rules (17), (18) and (19a/b) accounts for the coordination of sentences in which the complementation is intransitive. Rule (19a) describes the coordination of full conjoins, whereas rule (19b) describes instances of gapping, where the predicate in the second conjoin must be assumed to be missing. Rewrite rules for sentence include alternatives describing 'incomplete' sentences. For example, for intransitives we have the following rewrite rule:

(rewrite rule 20)

```
SENTENCE ("INTRANSITIVE", "MISSING PREDICATE"):
   CORE ("", "MISSING VERB").
```

This rule can be applied in order to account for the second, reduced conjoin in a coordination of sentences with intransitive complementation. For example, the analysis of a sentence like

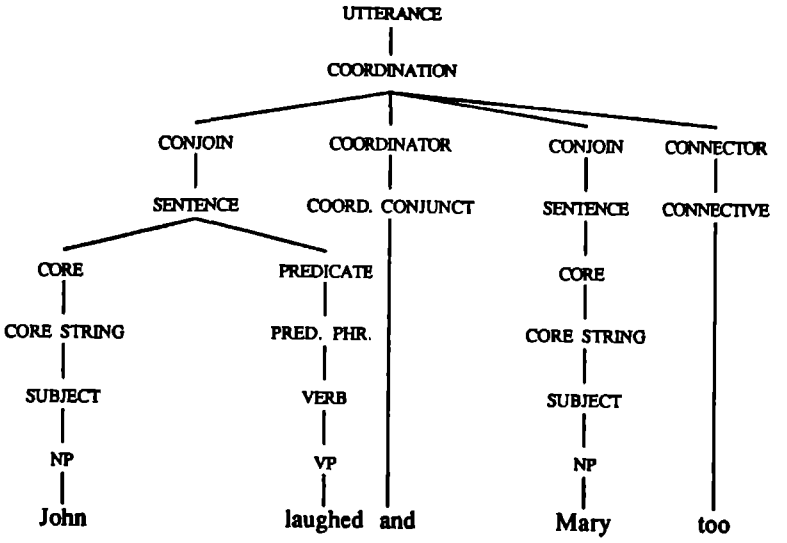
(30) John laughed and Mary too.

can be represented by means of the tree diagram below (Figure 22).

The second alternative of rule (16) in combination with rules (17), (18) and (19c) accounts for instances of coordination and also possibly for instances of gapping. The complementation must not be intransitive and must be the same in both conjoins. We have an instance of full coordination in case each of the conjoins has a "FULL" realization (as well as the same complementation). In case we come across another value for the realization of the second conjoin we have an instance of gapping. Other values for 'realization' include "MISSING VERB", "MISSING OD", "MISSING OI", and "MISSING OI OD".[11]

[11] As may be inferred from Table 1 this list is not exhaustive.

Figure 22



Consider the following examples:

(31) He painted the doors black and she the window-sills.
(32) Malcolm gave her a book and she as well.

In order to achieve correct analyses for the elements that are present in the second conjoins of the coordinations we must assume that in (31) the verb and the object complement are missing, while in (32) this is the case for the verb, the indirect object and the direct object. The derivation trees for (31) and (32) are given in Figures 23 and 24 respectively.

The description of instances of full coordination in which the complementation of the conjoins differs has been given in rules (16), (17), (18) and (19d).

Figure 23

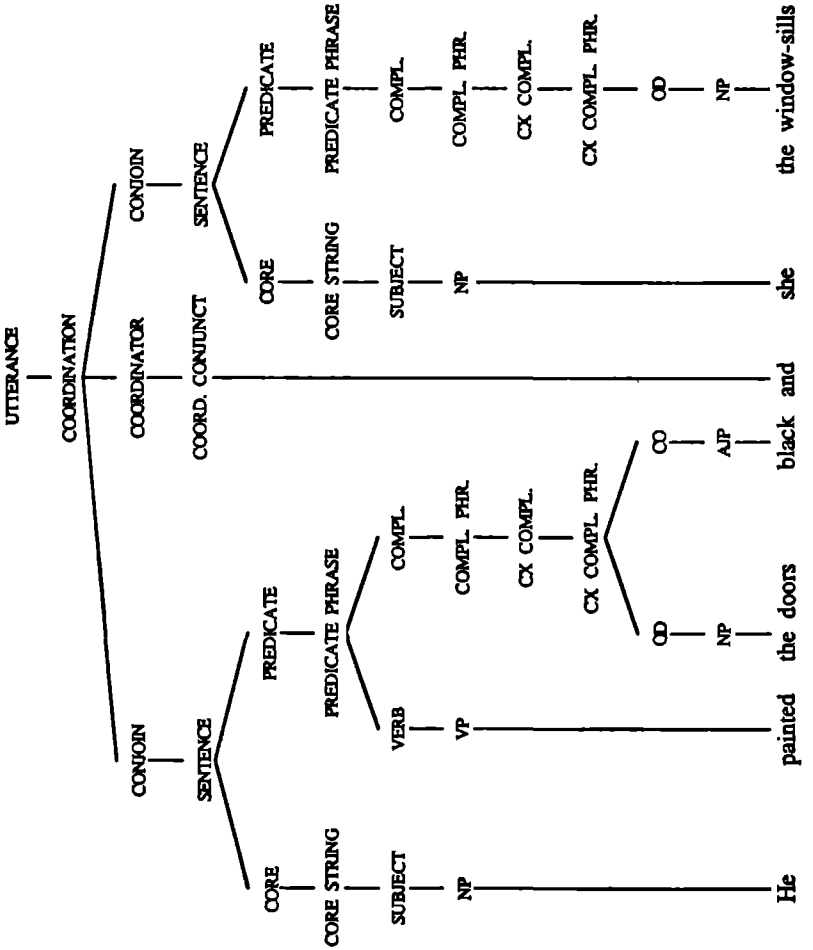
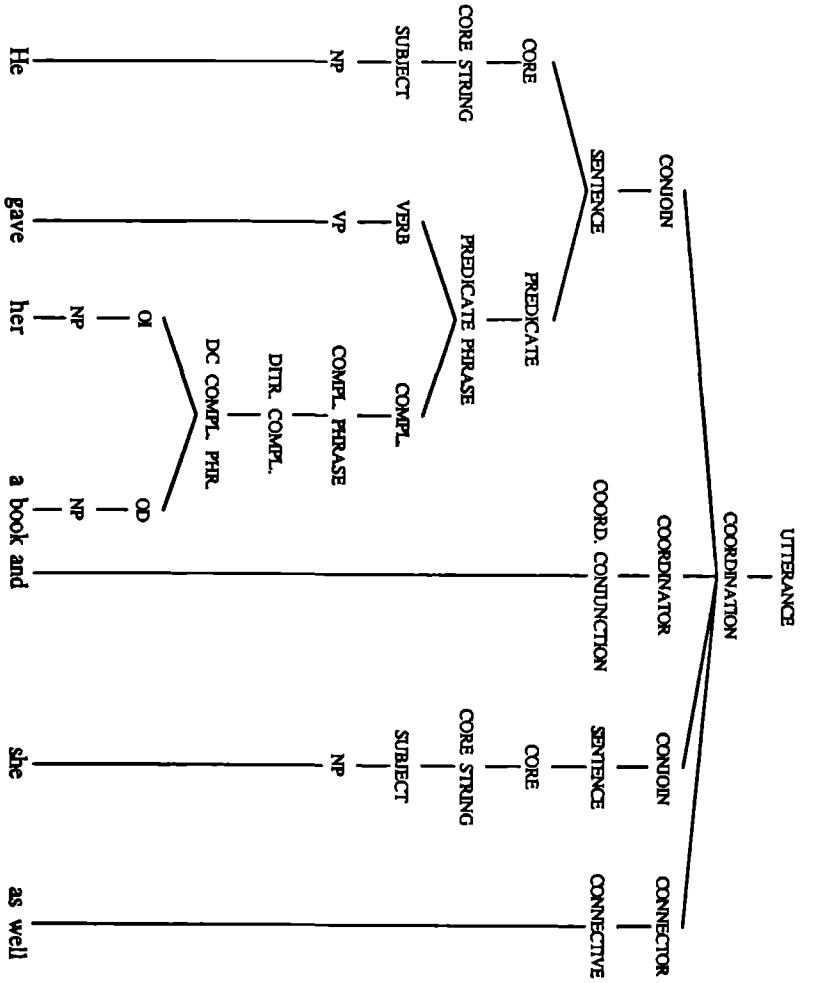

UTTERANCE
COORDINATION
CONJOIN
SENTENCE
CORE
CORE STRING
SUBJECT
NP
He
PREDICATE
PREDICATE PHRASE
VERB
painted
COMPL.
COMPL. PHR.
VP
CX COMPL.
CX COMPL. PHR.
NP
the doors
COORDINATOR
COORD. CONJUNCT
CONJOIN
AJP
black
and
SENTENCE
CORE
CORE STRING
SUBJECT
NP
she
PREDICATE
PREDICATE PHRASE
COMPL.
COMPL. PHR.
CX COMPL.
CX COMPL. PHR.
NP
the window-sills

Figure 24


UTTERANCE
COORDINATION
CONJOIN
SENTENCE
CORE
CORE STRING
SUBJECT
NP
He
PREDICATE
PREDICATE PHRASE
VERB
VP
gave
COMPL.
COMPL. PHRASE
DITR. COMPL.
DC COMPL. PHR.
OI
NP
her
OD
NP
a book
COORDINATOR
and
COORD. CONJUNCTION
CONJOIN
SENTENCE
CORE
CORE STRING
SUBJECT
NP
she
CONNECTOR
CONNECTIVE
as well

4.3 The noun phrase

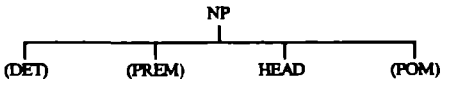
In this section, by way of illustration, an account is given of one of the modules of the grammar, that is, we describe the development of the formalized rules of the grammar that account for the noun phrase, concentrating on the problem of integrating the coordination module. Proceeding from an outline of the affix-free basis, the integration of the coordination module is discussed. Next the introduction of a number of affixes is discussed in some detail. Attention is also given to what generally are found to be more problematic aspects of the description of the NP, such as the incorporation of limiters and deferred determiners, and the description of apposition. We only briefly touch upon some of the descriptive problems that we come across when considering the NP as part of a larger constituent.

The basic NP structure

The description of the English NP in terms of an EAG calls for an outline of the affix-free basis to begin with. As observed earlier, our approach is basically similar to that of Quirk et al. (1985, 1972), and Aarts and Aarts (1982) in that we have opted for a description in terms of immediate constituents, incorporating both functions and categories.

In the structure assigned to the NP by for instance Quirk et al., four function slots are distinguished, namely those of determiner (DET), premodifier (PREM), head (HEAD), and postmodifier (POM). Apart from the function of head, all functions within the NP are optional. This structure can be represented as follows:

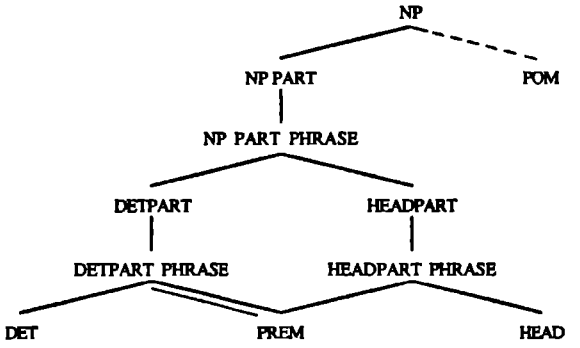
Figure 25: Flat NP structure



However, on the basis of the arguments provided in sections 3.5 and 4.2, the flat NP structure was abandoned and replaced by a somewhat different structure, one that is multi-layered. This structure, illustrated in Figure 26, still accommodates the four function slots we had before,

but includes another three, that of NP PART, DETPART and HEADPART.[12] The introduction of the additional function slots helps us overcome the descriptive problems that arise with respect to the description of instances of coordination that do not involve 'full' (traditional) constituents.

Figure 26: Multi-layered NP structure



By assigning this structure to the NP it becomes possible to base the grammar on elements that are actually present. Thus we are able to describe the coordination of strings like

(33) ten naïve and ten experienced marihuana smokers
(34) a successful writer and wonderful actor
(35) this old man and his wife from Bath

[12] There exists a variant of (part of) the structure presented here that allows us to deal with instances of shifted premodification as found in

too difficult a problem to solve

where the premodifier precedes rather than follows the determiner. In the variant structure both premodifier and determiner are obligatory constituents. Consequently, the headpart phrase then dominates a single function, the head.

as the coordination of detpart phrases (33), headpart phrases (34) and np part phrases (35). In the 'flat' approach this would have to be regarded as the coordination of NPs with ellipsis of the head in the first conjoin (33), the coordination of NPs with ellipsis of the determiner in the second conjoin (34), and the coordination of NPs with ellipsis of the postmodifier in the first conjoin (35).[13]

We can now formulate the following rules describing the affix-free NP structure:[14]

NP:
NP PART,
POM OPTION.

---

[13] In the approach of Quirk et al. it remains unclear whether coordination is assumed to involve functions or categories: apart from the coordination of noun phrases we may have coordination of pre- or postmodifiers, and of heads. Thus according to Quirk et al. (1972: 597-607) we have

- coordination of NPs in

  Old (men) and young men were invited.
  I don't care whether he is a studious or lazy undergraduate.

- coordination of pre- or postmodifiers in

  Honest and clever students always succeed.
  The bus for the Houses of Parliament and Westminster Abbey will soon be here.

- coordination of heads in

  Black boys and girls filled the classroom.
  Old books and magazines were given to the children to play with.

As we pointed out in section 4.2, we assume coordination to involve categories dominated by one and the same function node.

[14] A full detpart phrase, i.e. one that dominates a determiner followed by a premodifier, is only found in coordinations of detpart phrases (cf. example 33). In all other instances the detpart phrase will merely dominate the determiner, while any premodifier is dominated by the headpart phrase.

```
NP PART:
    NP PART PHRASE.

NP PART PHRASE:
    DETPART OPTION,
       HEADPART.

HEADPART:
    HEADPART PHRASE.

HEADPART PHRASE:
    PREM OPTION,
       HEAD.

DETPART OPTION:
      ;
    DETPART.

DETPART:
    DETPART PHRASE.

DETPART PHRASE:
    DET,
       PREM OPTION.

POM OPTION  : ; POM.
PREM OPTION : ; PREM.
```

The integration of the coordination module

Since each of the function nodes can dominate a coordination of categories we can reformulate our rules so that they will call upon the rules for 'lower level' coordination as contained in the coordination module. As was observed in section 4.2 (note 7) the set of coordination rules addressed here is basically similar to that for 'higher level' coordination and therefore only in part presented here.[15]

[15] For the time being we do not concern ourselves with the realization of the determiner and the modifiers, while the realization of the head is restricted to that by a common noun, proper noun, or pronoun.

```
NP:
    NP PART,
        POM OPTION.

NP PART:
    NP PART PHRASE;
    LOWER LEVEL COORDINATION ("NP PART").

NP PART PHRASE:
    DETPART OPTION,
        HEADPART.

HEADPART:
    HEADPART PHRASE;
    LOWER LEVEL COORDINATION ("HEADPART").

HEADPART PHRASE:
    PREM OPTION,
        HEAD.

DETPART OPTION:
      ;
    DETPART.

DETPART:
    DETPART PHRASE;
    LOWER LEVEL COORDINATION ("DETPART").

DETPART PHRASE:
    DET,
        PREM OPTION.

HEAD:
    COMMON NOUN;
    PROPER NOUN;
    PRONOUN;
    LOWER LEVEL COORDINATION ("HEAD").

POM OPTION  : ; POM.
PREM OPTION : ; PREM.

LOWER LEVEL COORDINATION (function + poss info):
    LOWER LEVEL SYNDETIC COORDINATION (function + poss info);
    CORRELATIVE COORDINATION (function + poss info),
        restricted to particular functions only (function).
```

```
LOWER LEVEL SYNDETIC COORDINATION (function + poss info):
    CONJOIN (category 1, function + poss info 1),
        CONSTITUENT SEQUENCE (category 2, function + poss info 2),
            identical or different categories (category 1, category 2, function),
                any additional info (function, poss info 1, poss info 2,
                                                                poss info).

CONSTITUENT SEQUENCE (category, function + poss info):
    COORDINATOR (type),
        CONJOIN (category 1, function + poss info 1),
            MORE CONSTITUENTS (category 2, function + poss info 2),
                identical or different categories (category 1, category 2, function),
                    any additional info (function, poss info 1, poss info 2,
                                                                poss info).

CORRELATIVE COORDINATION (function + poss info):
    COORDINATOR (type),
        CONJOIN (category 1, function + poss info 1),
            COORDINATOR (type),
                CONJOIN (category 2, function + poss info 2),
                    identical or different categories (category 1, category 2,
                                                                function),
                        any additional info (function, poss info 1, poss info 2,
                                                                poss info).

restricted to particular functions only ("SU"): .
restricted to particular functions only ("NP PART"): .
```

As before the first predicate in the rules describing coordination (identical or different categories ...) allows for identical categories to be coordinated in case they are dominated by one and the same function node, by way of rewrite rule (4) (repeated here):

```
identical or different categories (category, category, function): .
```

For the NP this rule will apply to instances where NP PART, HEADPART, DETPART, and HEAD are found to have a multiple realization through coordination.

Since coordination is assumed to involve categories dominated by one and the same function node we must have a function node dominating the NP in order to make it possible to have coordination of NPs.

Thus if we take for example the function of subject to be dominating the NP, we can account for both single NPs but also for coordinations involving NPs by way of the following rule:

```
SUBJECT:
    NP;
    LOWER LEVEL COORDINATION ("SU").
```

where the second alternative calls upon the coordination module. Here too rewrite rule (4) applies. Note, however, that unlike the NP PART PHRASE, HEADPART PHRASE, DETPART PHRASE, and the categories realizing the function of HEAD, the NP may be involved in a coordination with one or more different categories, in which case rewrite rule (4) no longer applies but rather a rule with the format

```
identical or different categories (category 1, category 2, function):
    not equal (category 1, category 2).
```

where the literal affix values for 'category 1' and 'category 2' are assumed to be not equal.

The second predicate rule (any additional info ...) will not be discussed here. Such a discussion would require that the other affixes that are used in the description of the NP had been discussed in some detail. Since at this point this is not the case, for the moment the application of a rewrite rule with the format

```
any additional info (function, "", "", ""): .
```

is presumed.

In the rules above it is possible to have coordination at four different levels in the NP. Thus it is possible to have[16]

[16] In the derivation trees below we have once more 'filtered out' the intermediate labelling of constituents in coordinations, such as CONSTITUENT SEQUENCE and MORE

- a coordination dominated by the function of HEAD; e.g. a string like

  (36) all these men and women interested in Chinese painting

  where the postmodifier is assumed to modify both *men* and *women* (in which case also the determiner must be associated with both these nouns) yields the analysis represented in Figure 27. The alternative interpretation of this string results in an analysis (not presented here) in which the coordination is taken to be a coordination of NPs.

- coordination of HEADPART PHRASEs; for example

  (37) this loyal friend and trusted partner of yours

  A representation of the analysis is given in Figure 28.

- coordination of DETPART PHRASEs; for example

  (38) ten naïve and ten experienced marihuana smokers

  See Figure 29.

- coordination of NP PART PHRASEs; for example

  (39) both the old man and his youngest daughter from Bath

  A representation of the analysis is given in Figure 30.

- coordination of NPs; for example

  (40) the small elephant you see over here and that one over there, which were both born in captivity

  will be analyzed as represented in Figure 31.

CONSTITUENTS.

Figure 27

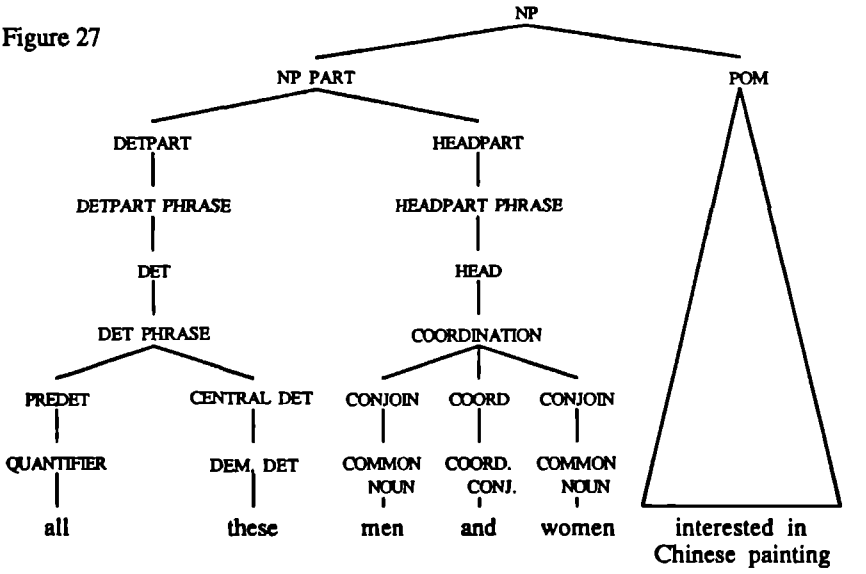


Figure 28

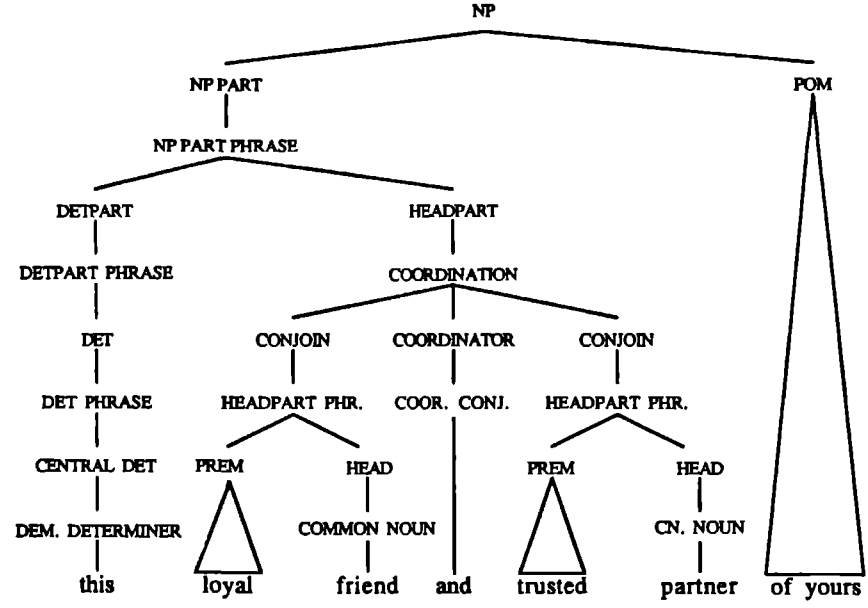
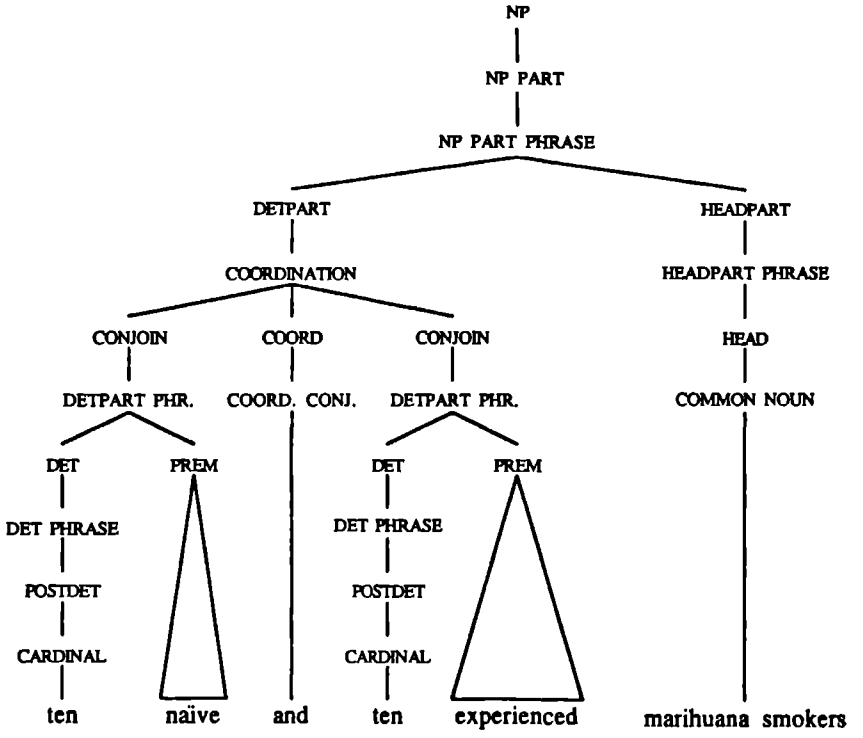


Figure 29

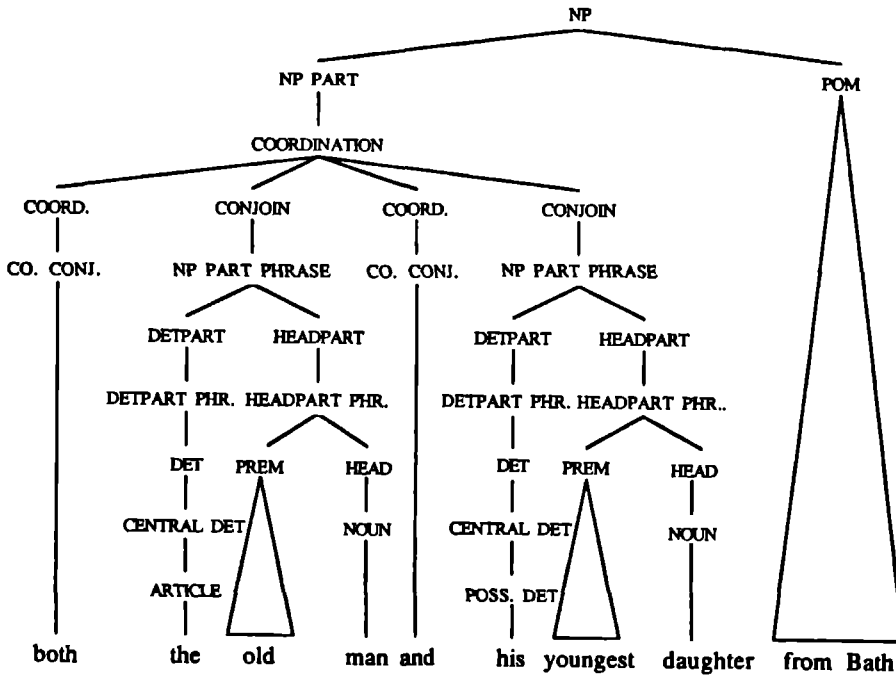


Note that correlative coordination only occurs at the two highest levels, i.e. the level of NP and NP PART PHRASE. This restriction is expressed in terms of the predicate 'restricted to particular functions only (function)'. Also, the highest level postmodifier (below referred to as POM 1) will be recognized only if it follows a coordination of NPs; in all other instances any postmodifier is recognized as a postmodifier at the same level as NP PART (below referred to as POM 2), i.e., it is assumed to be dominated by the NP node.

Having formulated the rules describing the structure of the NP as discussed above, we allow for the ambiguous analysis of strings answering (in part) to the pattern

(COORD)-(DET)-(PREM)-HEAD-(POM2)-COORD-(DET)-(PREM)-HEAD-(POM2)-(POM1)

Figure 30



in case both POM 2s are absent or in case the first POM 2 and POM 1 are absent. For example, if we look once more at example (39), it appears that two analyses are possible:

(a) (both (the old man) and (his youngest daughter from Bath))
POM 2

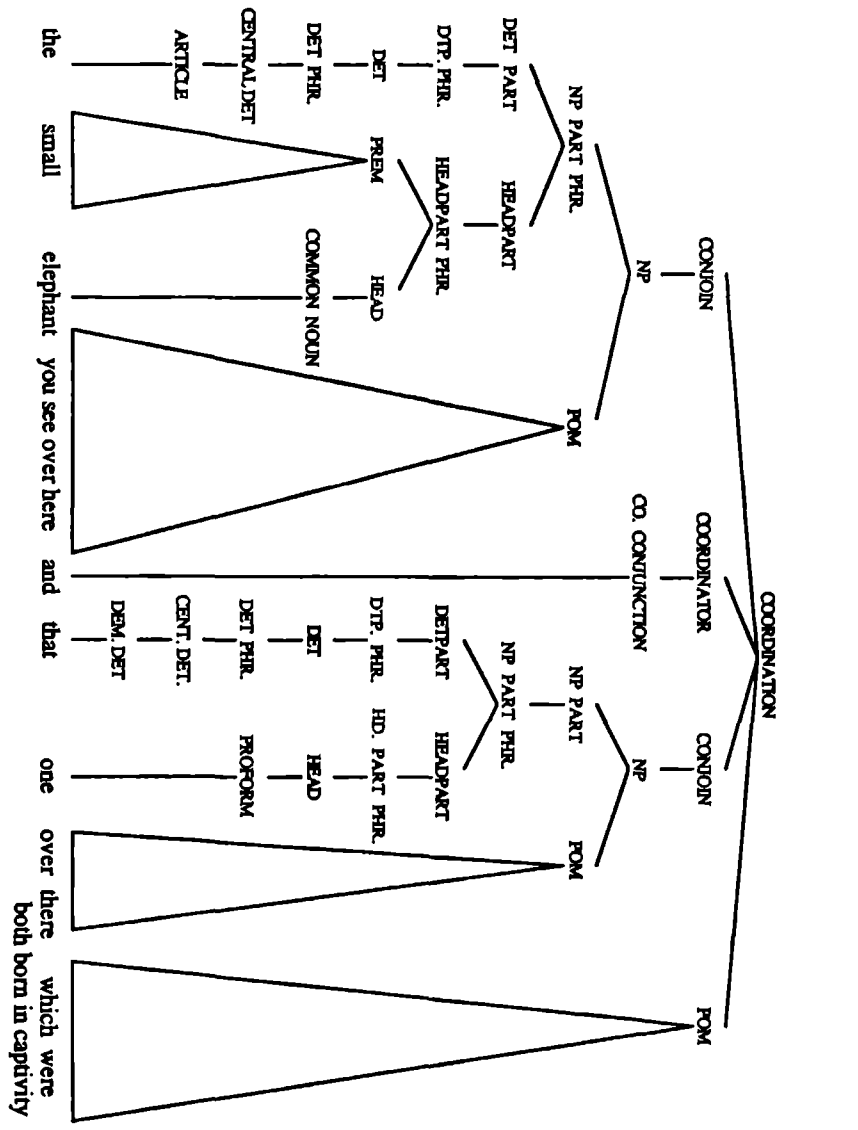
(b) (both (the old man) and (his youngest daughter) from Bath)
POM 1

The same goes for

all these men and women interested in Chinese painting[17]

[17] We regard *all these* as one determiner constituent in view of the fact that an NP like *all these*

Figure 31


COORDINATION
CONJOIN
COORDINATOR
CONJOIN
POM
NP
CO. CONJUNCTION
NP
POM
NP PART PHR.
NP PART
NP PART PHR.
DET PART
HEADPART
DETPART
HEADPART
DTP. PHR.
HEADPART PHR.
DTP. PHR.
HD. PART PHR.
DET
PREM
DET
HEAD
DET PHR.
COMMON NOUN
DET PHR.
PROFORM
CENTRAL DET
CENT. DET.
ARTICLE
DEM. DET
the small elephant you see over here and that one over there which were both born in captivity

which in addition to the analysis given in Figure 27 may receive the analyses

(a) (all these men) and (women interested in Chinese painting)
POM 2
(b) (((all these men) and (women)) interested in Chinese painting)
POM 1

Affixes in the NP grammar

Having established the basic structure for the NP we can now introduce a number of affixes which we feel are relevant either within the NP itself or when the NP occurs as part of a larger constituent. They are briefly commented upon below.

'detstructure' and 'headreal'

The affixes 'detstructure' and 'headreal' are introduced to indicate which determiners are permitted or even required with particular headrealizations. In the rules presented above the realization of the head was restricted to that by a common noun, proper noun, or pronoun, while so far we have not concerned ourselves with the realization of the determiner. Here we first consider the internal structure of the determiner and some of its realizations. Then the relation between determiner structure and realization of the head is discussed in some detail. We conclude this subsection by giving the formal rules that describe the observed dependencies.

In its simplest form the function of determiner in a noun phrase is realized by a single item. For example,

(41) *all* people present
(42) *the* children and *their* parents

---

men and women may have the interpretations (1) and (2) but not (3) or (4):

(1) (all these men) and (women)
(2) (all these men) and (all these women)
(3) (all these men) and (all women)
(4) (all these men) and (these women)

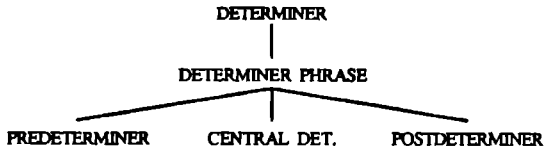
(43) *five* oranges

In a great number of instances, however, we find more than one determiner item:

(44) *the first eight* agreements
(45) *the third* argument
(46) *such a* categorical approach

When this is the case, the order in which these determiners occur is not free. Rather, three subclasses of determiners must be distinguished according to the relative positions they can be found in. This leads us to introduce a DETERMINER PHRASE as the unique realization for the function of determiner, while within this phrase three function slots are distinguished. A representation of the structure of the determiner can be found in Figure 32.

Figure 32: Determiner structure



The first function slot may be occupied by items that can co-occur with articles, demonstratives and suchlike items, in which case they precede these. The first function slot in the DETERMINER PHRASE is, therefore, referred to as the PREDETERMINER. Items that typically occur as predeterminer are the quantifiers *all, both, half, such,* and *many.* They are mutually exclusive.

In the function of CENTRAL DETERMINER we find articles (47), assertive (48) and non-assertive determiners (49), demonstratives (50), relative determiners (51), interrogative determiners (52), negative (53) and possessive determiners (54), and genitive noun phrases (55).

(47) *a* row
(48) *some* other Allied military outfit
(49) *any* red flags
(50) *this* time
(51) Appleby, *whose* instincts were always humane
(52) *what* newspaper

(53) *no* overriding reason
(54) *your* friend
(55) *the Director General's* arguments

Central determiner items are mutually exclusive.

The function of POSTDETERMINER is realized by a cardinal or ordinal numeral (56)-(57) or a quantifier (58).

(56) *three* nights
(57) the *first* case
(58) *several* industrialized countries

Unlike predeterminer and central determiner items, some postdeterminers can co-occur. For example,

(59) the *first few* weeks
(60) the *last two* motions

As we observed above, a determiner may be realized by a single item. With the introduction of the notion of determiner phrase and the internal structure that is associated with it, realization by a single item must be interpreted as the realization of one of the three functions we distinguished. In other words, we find a minimal realization of the determiner if the PREDETERMINER is realized (41), or the CENTRAL DETERMINER (42), or the POSTDETERMINER (43). In case the determiner is realized by more than one item, it appears that any of the following combinations is possible:

- a predeterminer followed by a central determiner; for example,

  (61) *such a* time
  (62) *both my* children

- a predeterminer followed by one or more postdeterminers; for instance,

  (63) *all four* books
  (64) *many other first* experiences

- a predeterminer followed by a central determiner followed by one or more postdeterminers; consider the following examples:

(65) *all these six* children
(66) *both their last two* ratings

- a central determiner followed by one or more postdeterminers, as in

(67) *the next forty* miles
(68) *the same three* men

- two or more postdeterminers; for example,

(69) *another two* exams
(70) *most other* people

When we formalize the above, the following context-free rules result:

```
DET: DET PHRASE.

DET PHRASE:
    PREDET,
        CENTRAL DET OPTION,
            POSTDET OPTION;
    CENTRAL DET,
        POSTDET OPTION;
    POSTDET,
        POSTDET OPTION.

CENTRAL DET OPTION:
        ;
    CENTRAL DET.

POSTDET OPTION:
        ;
    POSTDET,
        POSTDET OPTION.
```

With the help of affixes which record which function(s) has/have been realized, the above rules can be converted to the EAG rules below. Here the affix 'pre' can take the values "PRE" or "EMPTY", the affix 'central' the values "CENTRAL" or "EMPTY", and 'post' the values "POST" or "EMPTY". The predicate 'minimal realization ...' requires

that at least one of the functions of the determiner phrase is realized.[18]

```
DET : DET PHRASE.

DET PHRASE :
    PREDET OPTION (pre),
        CENTRAL DET OPTION (central),
            POSTDET OPTION (post),
                minimal realization (pre, central, post).

PREDET OPTION ("EMPTY") : .
PREDET OPTION ("PRE")   : PREDET.

CENTRAL DET OPTION ("EMPTY")   : .
CENTRAL DET OPTION ("CENTRAL"): CENTRAL DET.

POSTDET OPTION ("EMPTY") : .
POSTDET OPTION ("POST") :
    POSTDET,
        POSTDET OPTION (post).

minimal realization ("PRE", central, post) : .
minimal realization ("EMPTY", "CENTRAL", post) : .
minimal realization ("EMPTY", "EMPTY", "POST") : .
```

In our rules so far, we have assumed that noun phrases are introduced by a determiner. The determiner must, however, be considered to be an optional constituent, its optionality being conditioned by the nature of the head. For example, with common nouns the NP is generally (although not necessarily) introduced by a determiner, whereas with proper nouns and personal pronouns there is no determiner. Moreover, not in all instances where a determiner may be found is it possible to have a determiner phrase with all its functions realized. The following restrictions can be observed:

- if the head of the NP is realized by a nominalized predeterminer quantifier such as *all* or *both*, by any kind of pronoun other than

---

[18] The empty realization of the determiner was already accounted for in the rule

```
DETPART OPTION : ; DETPART.
```

demonstrative, or by a proper noun, no determination of the NP is possible;

- if the head is realized by a demonstrative pronoun, the determiner phrase consists merely of a predeterminer; for example,

  (71) all this
  (72) both these

- only when the head is realized by a common noun, one of the proforms *one* or *ones*, a cardinal or ordinal numeral, or a postdeterminer quantifier, is it possible to have a determiner phrase in which any one of the functions (PREDET, CENTRAL DET and POSTDET) may be realized. Consider the following examples:

  (73) *twice this much* money
  (74) *all her four* children
  (75) *all these countless* hours
  (76) *half the other* one
  (77) *both their two* last
  (78) *all the three* same

In order to distinguish different head realizations, we introduce the affix 'headreal'. This affix, which is associated with the head, records the realization of the head and relates this information to the determiner. The meta-rule for 'headreal' is the following:[19]

headreal :: "PRE Q"; "PN"; "PRN"; "DEM"; "CN"; "PRO"; "NUM"; "POST Q".

The affix information that is associated with the pre-, central, and postdeterminer is combined in the affix variable 'detstructure' so that the rules that were given above for the determiner and the determiner

[19] The abbreviations that are introduced here are to be interpreted as follows:

PRE Q	predeterminer quantifier	CN	common noun
PN	pronoun	PRO	proform
PRN	proper noun	NUM	numeral
DEM	demonstrative pronoun	POST Q	postdeterminer quantifier

phrase then look as follows (the other rules remain the same):

```
DET (detstructure) : DET PHRASE (detstructure).

DET PHRASE (pre + central + post) :
    PREDET OPTION (pre),
        CENTRAL DET OPTION (central),
            POSTDET OPTION (post),
                minimal realization (pre, central, post).
```

The relation between the affixes 'detstructure' and 'headreal' is made explicit by means of a predicate 'detstructure depends on headreal ...'. The formal representation of the observed restrictions can be found in the following rules:

```
detstructure depends on headreal ("EMPTY", "PRE Q") : .
detstructure depends on headreal ("EMPTY", "PN") : .
detstructure depends on headreal ("EMPTY", "PRN") : .
detstructure depends on headreal ("PRE", "DEM") : .
detstructure depends on headreal ("EMPTY", "DEM") : .
detstructure depends on headreal (detstructure, "CN") : .
detstructure depends on headreal (detstructure, "PRO") : .
detstructure depends on headreal (detstructure, "NUM") : .
detstructure depends on headreal (detstructure, "POST Q") : .
```

'countability'

By means of an affix 'countability' we describe the "countability relation" that exists between the determiner and the head. The determiners carry the value "SING", "PLU", or "MASS", depending on the countability of the heads they occur with.

'number' and 'person'

In order to be able to describe certain relations on sentence-level, such as subject-verb concord, the affixes 'number' and 'person' are introduced.

The affix 'number' can have the values "SING" or "PLU". Note that, although in a large number of cases the value associated with 'number' will be the same as the value associated with 'countability', there is an important difference between these two affixes: whereas 'countability' is used in describing the relation between the determiner and the head, 'number' functions on sentence-level. Since it would make no sense to speak of "MASS" when describing relations like subject-verb concord, 'number' only has the values "SING" and "PLU". The need for a separate affix 'number' also arises from the fact that, since we allow coordination at various levels, we have to remember what the countability of the head was, until the head (and of course the premodification, if there is any) is joined with the determiner; whereas the value for 'countability' remains the same, even when coordination takes place, the value for 'number' is possibly subject to change.

The affix 'person' can have the values "1ST", "2ND" or "3RD". In NPs without any coordination the value for 'person' will be "3RD", except for those NPs where the head is realized by a personal pronoun. Like 'number' the value for 'person' may change under the influence of coordination.

Testing the grammar

The writing of the initial version of the NP grammar with the inclusion of the affixes we described above was a purely theoretical affair. At the time the first version was written it was impossible to obtain any feedback from immediate testing due to the fact that the parser generator had not yet fully been developed. When we came to test the NP grammar, initial testing of the NP in isolation on a small test corpus that had been compiled for the purpose yielded quite satisfactory results. However, already at this point it was clear that ambiguity arose easily. The multi-layered structure assigned to the NP needed to be controlled very carefully in order to keep this ambiguity within limits.[20] Originally in writing the rules for coordination it had been assumed that coordination occurred at any level where conditions with respect to the realization of the conjoins and associated features were met. This proved too weak a restriction. Therefore, a predicate rule was introduced as a kind of con-

[20] At a later stage the same problem was observed in the description of the adjective phrase and the adverb phrase.

trol mechanism, which would check whether a particular node was single or multiple branching. In general, coordination is only permitted with multiple branching nodes. There are a few exceptions to this rule. One exception may be found in a node like HEAD, which is always single branching. The multiple-branching condition is also superseded at the level of NP, where, dominated by single branching nodes, NPs may be assumed to be coordinated if the coordination involves different (lexical) categories. With these provisions we managed to avoid the potential fourfold ambiguous analysis of a string like

(79) Peter and I

which otherwise might have been analyzed as the coordination of a PROPER NOUN and a PRONOUN, HEADPART PHRASEs, NP PART PHRASEs, or NPs.[21]

The NP grammar proved to be quite satisfactory when used for the analysis of NPs in isolation. As one of the modules in the sentence grammar, however, it fell short in various respects. Contrary to our expectations, these shortcomings did not so much concern the 'incompleteness' of our description[22]; rather, it was the strict regulation of particular relationships that caused the analysis to fail in a number of instances. For example, the description of the determiner-head relationship was rather strict in its definition of what determiner structure occurred with what head realization. Although on the whole the description was adequate, it failed to take into account the fact that there are exceptions to the rule. Among the instances that the grammar failed to analyze were those where there was a generic use of a (singular) noun (80) or where a proper noun was used as a common noun (81). Moreover, failure of analysis occasionally occurred with noun phrases that in their function of prepositional complement in a preposi-

---

[21] Under the conditions stated above a string like

Peter and I

will be analyzed as the coordination of NPs, the coordination of different lexical categories being prohibited at the level of HEAD, HEADPART and NP PART.

[22] We had been aware of the fact that a description of genitive noun phrases, appositives, and nominal adjectives as heads of NPs was lacking.

tional phrase lacked the determiner one would normally expect (82).

(80) *Man* is a threat to many forms of wildlife.
(81) *This John I spoke to the other day* seemed quite happy about it.
(82) The children stayed at *school.*

The observation that these cases generally involve a special usage -- generic, more or less idiomatic -- is of little avail: instances like these occur frequently and their analysis should not present any problems. On the basis of the test results it was therefore decided to refrain from a description of the determiner-head dependency (expressed earlier in the predicate 'detstructure depends on headreal'). Further testing indicated that this had the desired effect: the analysis was achieved of NPs that before could not be analysed as a result of the restrictions which had been imposed on the determiner and the head. Whereas the affix 'detstructure' became obsolete, the affix 'headreal' retained its usefulness in the description of coordination.

Testing the NP grammar as a module in the sentence grammar, we also found that the analysis of coordinations involving coordinating conjunctions like *or, nor,* and the correlatives *either ... or* and *neither ... nor* was not unproblematic. This brought up the question to what extent the description of subject-verb concord was really useful in the analysis of corpus material. In case the conjoins in such a coordination agree with one another with respect to number and person, the coordination simply carries the same values. However, in coordinations where the conjoins differ from each other in number and/or person it appears impossible to determine what the value(s) should be. Consider the following example

(83) Either you or your friend made a mistake.

Although speakers will generally avoid such a sentence when the verb is in the perfect or progressive and therefore requires a finite auxiliary showing number and person, a sentence like (83), with the verb in the past tense, appears perfectly acceptable. Aiming at a description of subject-verb concord we are faced with the fact that such sentences remain ambiguous in their analysis: for instance, in the case of our example (83) there will be three analyses for *made*, namely 'second person singular', 'second person plural' and 'third person singular'. This and the fact that we do come across instances where there is no grammatical subject-verb concord, set against the observation that only few

analyses profit from the description of subject-verb concord (by way of a reduction in ambiguity), have led us to conclude that such a description had better be left out.

The genitive noun phrase

The description of the genitive noun phrase in terms of the EAG formalism was problematic because genitive noun phrases are by definition left-recursive.[23] Since the parser generator cannot handle left-recursion no parser will result. In order to overcome the problem we decided to mark the beginnings of genitive noun phrases, thus requiring intervention in the analysis process. Note, however, that as a side-effect the ambiguity that would otherwise arise in the analysis of strings like (84) and (85) no longer occurs.

(84) her husband's pet
(85) this men's wear

The disambiguating effect that the marking has is as follows: by placing a mark before the determiner 'her' in example (84), the genitive can only be interpreted by the parser as a specifying genitive; in (85) a mark placed after the determiner 'this' marks the genitive as a classifying genitive.

Apposition

In the first version of the NP grammar a description of apposition had been left out. The reasons for this were the following: first, handbooks on English grammar are not very specific where the description of apposition is concerned so that it is not easy to give an accurate description of this phenomenon; second, it was expected that the inclusion of a description of apposition would yield a large amount of (undesired) ambiguity even for rather simple noun phrases. This latter point

[23] The inherent left-recursiveness of the genitive noun phrase is strictly speaking restricted to specifying genitives since only these can occur both as determiners and as heads in noun phrases. With specifying genitives the determiner qualifies the genitive, not the head noun. Classifying genitives typically occur as premodifiers and may be preceded by a determiner qualifying the head noun.

can easily be demonstrated: assuming the description of apposition as an NP followed by a second NP without any restrictions whatsoever leads to the ambiguous analysis of strings like

(86) all this sugar

where we get four analyses:

(all this sugar)
((all) (this sugar))
((all this) (sugar))
((all) (this) (sugar))

The first analysis shows the analysis of the input string as a single NP. In the second analysis we find two NPs in apposition: *all* and *this sugar*. The third analysis takes *all this* and *sugar* to be appositive NPs. Finally, in the fourth analysis we find an apposition with three NPs: *all*, *this* and *sugar*.

Only at a later stage, after having come across various instances of apposition in the corpus, did we decide to attempt to include this phenomenon in our description. As we had already suspected, apposition was found to occur in many different forms and varying complexity. In the broadest sense apposition is the reformulation of one constituent by means of another. When given this interpretation, apposition comes to resemble coordination closely. Its description can take the form of rule-schemata which, when called upon, will operate according to the rule-generating rule principle. In a far more restricted interpretation of the notion, apposition is defined as two or more NPs between which there exists a reformulatory, specifying, or restricting relationship. In our grammar the latter interpretation of the apposition was adopted. Our grammar aims to describe such instances as

(87) tonight's movie, Guns from Navarone, starring Gregory Peck and David Niven
(88) two of my friends from highschool, Tom and Paul
(89) John, his brother and I, we

For the description of apposition in this restricted sense the following rules were formulated:

```
APPOSITION:
    APPOSITIVE,
        APPOSITIVE CONSTITUENT SEQUENCE.

APPOSITIVE CONSTITUENT SEQUENCE:
    ADVERBIAL OPTION,
        CONNECTOR OPTION,
            APPOSITIVE,
                ADVERBIAL OPTION,
                    CONNECTOR OPTION,
                        MORE APPOSITIVE CONSTITUENTS,
                            POM OPTION.

MORE APPOSITIVE CONSTITUENTS:
    ;
    APPOSITIVE CONSTITUENT SEQUENCE.

APPOSITIVE:
    NOUN PHRASE;
    COORDINATION("APP").
```

The rules basically describe APPOSITION as two or more APPOSITIVES. An APPOSITIVE is realized by a noun phrase or a coordination of NPs. Note that we allow postmodifiers to occur following an apposition of noun phrases. As with coordination, we assume that in noun phrases as found in (90) the postmodifier modifies all that precedes:

(90) this man, Mr Jones, chairman of the board of directors, who had only recently been appointed

The rules describing apposition also describe the possibility of connective items occurring as connectors in appositions. Among the items typically found in appositive constructions we find *that is, rather, for example,* etc.

The description of apposition in its restricted sense is not at all problematic. Unfortunately, however, it appears impossible to have the rules apply freely: they are apt to generate undesired ambiguity even with the simplest noun phrases.[24] Therefore, we decided that the analysis of

[24] The ambiguity arises from the fact that we describe not only apposition, but also nominaliza-

appositive noun phrases must await intervention by the linguist triggering the above rules. This being the case we turned this to our advantage and allow for 'floating appositives'. For example,

(91) They sidled awkwardly, six of them.
(92) Tomorrow we're going to be busy, you and I.

In instances like these a second or further appositive does not immediately follow an earlier appositive.

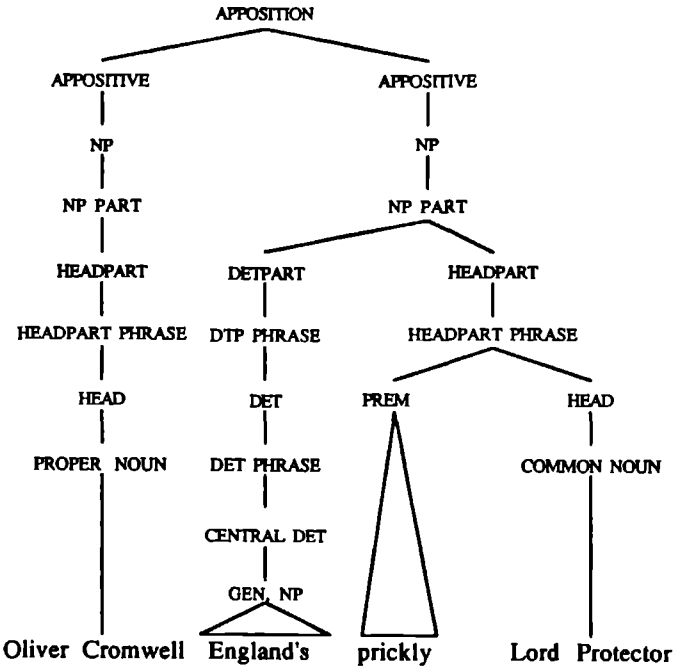
All in all, with the provision that we leave it to the linguist to trigger the rules for apposition (and genitive noun phrases), we find that the analysis of noun phrases like those found in (93)-(98) can be handled satisfactorily. The derivation trees for these NPs can be found in Figures (33)-(38).[25]

(93) Oliver Cromwell, England's prickly Lord Protector
(94) Disraeli's nemesis, Gladstone -- that unintentional instigator of the Bath Club Affair
(95) those base, servile, self-degraded wretches, Virgil and Horace
(96) John Ruskin, the fiercely moralistic essayist and art critic
(97) the geneticist L.L. Cavalli-Sforza, the anthropologist F.T. Clark and the ethologist J.M. Cullen
(98) James McNeill Whistler, the expatriate American artist and notable dandy, who was a champion of the new "art for art's sake" painters and writers

tion of determiners. This point was illustrated in example (86).

[25] As before, in the derivation trees we have filtered out the intermediate labelling of constituents in coordinations. Similarly, the intermediate labelling of constituents in appositions as APPOSITIVE CONSTITUENT SEQUENCE and MORE APPOSITIVE CONSTITUENTS has been left out.

Figure 33



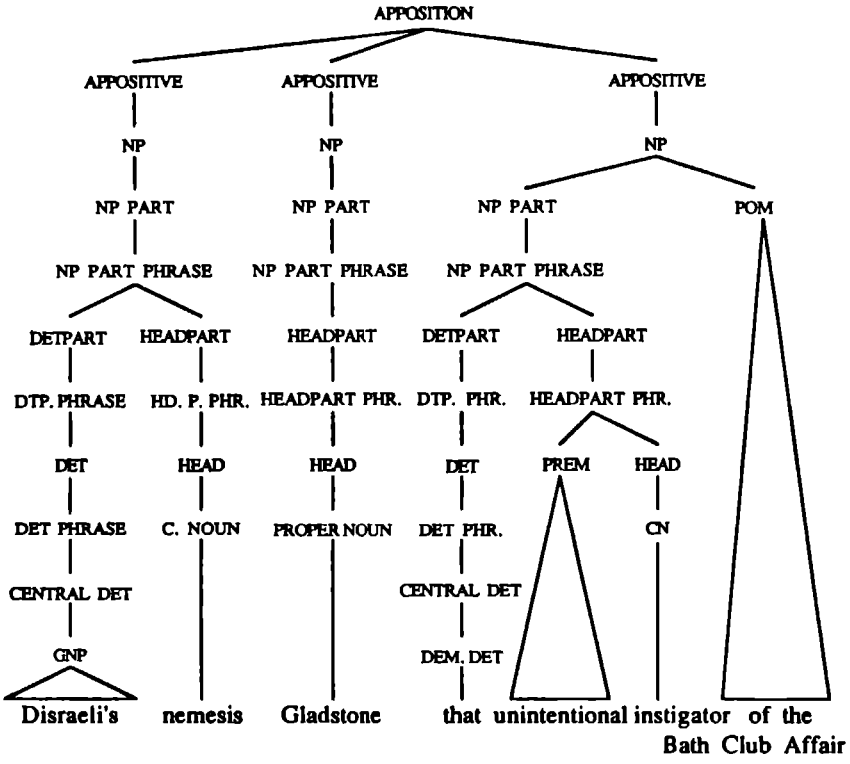
Focusing adjuncts

In describing the noun phrase obviously some provision has to be made in order to account for instances like those found in examples (99)-(103), taken from Quirk et al. (1972: 431).

(99) *Only the extremely wealthy customers* could afford to buy those.
(100) *At least ten workers* reported ill yesterday.
(101) *Especially the girls* objected to his manners.
(102) *The workers, in particular,* are dissatisfied with the government.
(103) *Even Bob* was there.

Items like *only, at least, especially, in particular, as well* and *even* in Quirk et al.'s definition are focusing adjuncts. They "make explicit

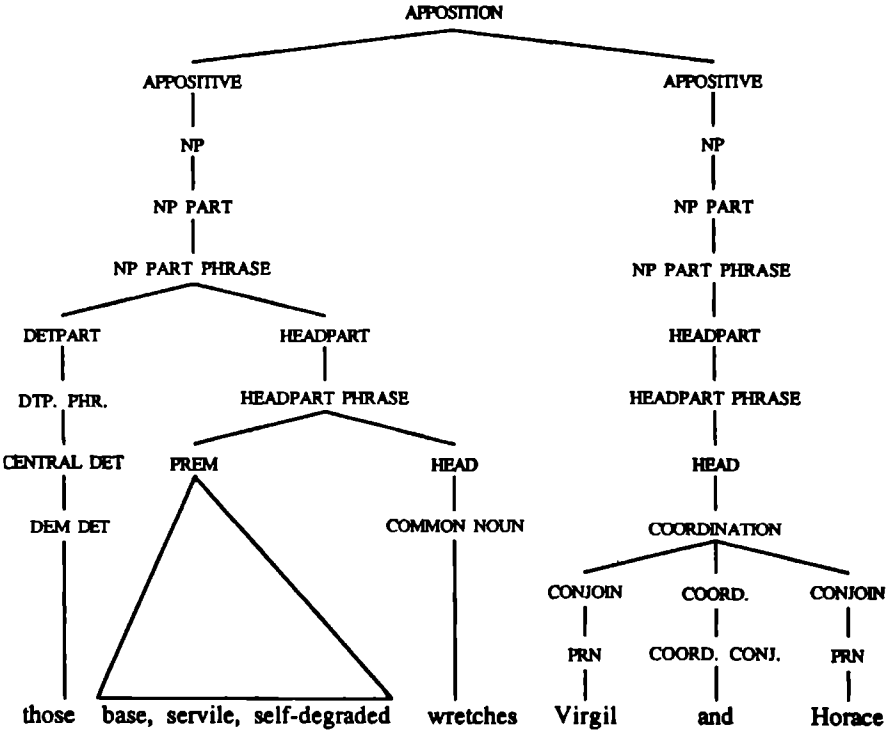
Figure 34



either that what is being communicated is restricted to a part that is focused, in which case they are called RESTRICTIVE ADJUNCTS, or that a focused part is an addition, in which case they are called ADDITIVE ADJUNCTS" (Quirk et al., 1972: 431). Quirk et al. argue that in instances such as those exemplified above focusing adjuncts should not be analyzed as part of the noun phrase but rather as an adverbial on clause or sentence level. Among the arguments they provide there is only one that we need discuss here, namely the fact that focusing adjuncts can focus on a noun phrase to which they are not juxtaposed. Analyzing focusing adjuncts as part of the noun phrase would be problematic in instances like (104) which could then only be regarded to be discontinuous.

(104) I don't want any beer, I *only* want *some water*.

Figure 35

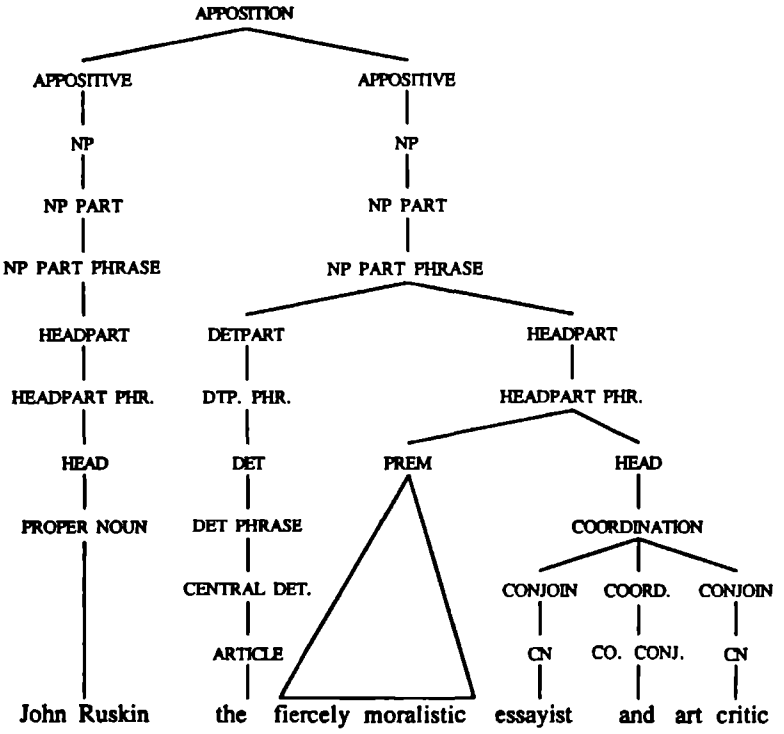


However, taking into consideration the syntactic features of adjuncts that Quirk et al. summarize earlier, it appears that focusing adjuncts behave rather differently from the other classes of adjuncts they distinguish (cf. Quirk et al., 1972: 426-429):

1. focusing adjuncts cannot come within the scope of clause negation and cannot be the focus of the negation; cf.

(105) They didn't see him *last night*, but they did see him *this morning*.
* (106) They didn't see him *at least* but they did see him *even*.
(107) They didn't see *only him* but *even Jack*.

Figure 36



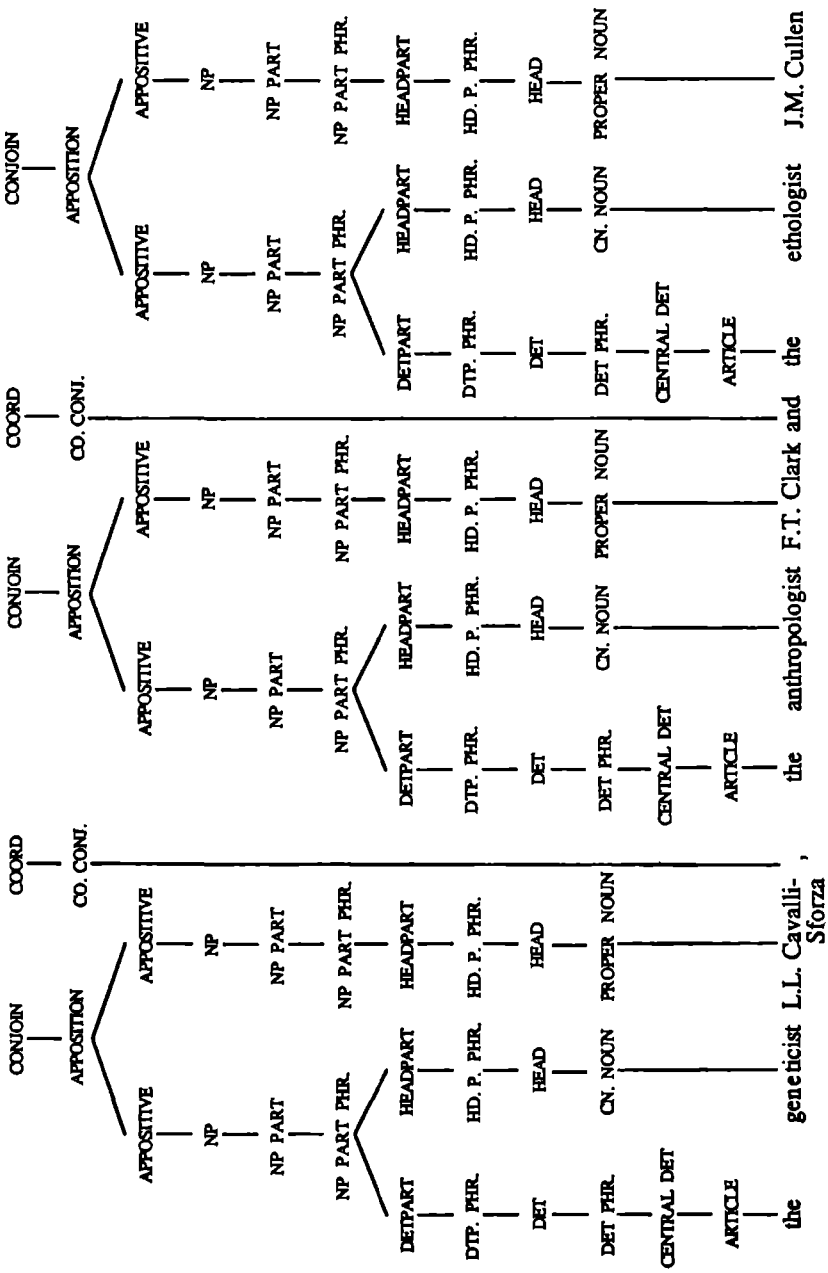
2. unlike other adjuncts, focusing adjuncts cannot be the focus of a cleft sentence; cf.

(108) It was *last week* that I visited John.
* (109) It was *even* that I visited John.

3. focusing adjuncts unlike other adjuncts cannot come within the scope of clause interrogation and cannot be the focus of the question; cf.

(110) Did you see him *yesterday* or did you see him *today*?
* (111) Did you see him *even*?
(112) Did you see *only him*?

e 37


CONJOIN
COORD
APPOSITION
CO. CONJ.
APPOSITIVE
APPOSITIVE
NP
NP
NP PART
NP PART
NP PART PHR.
NP PART PHR.
DETPART
HEADPART
DTP. PHR.
HD. P. PHR.
DET
HEAD
DET PHR.
CN. NOUN
PROPER NOUN
CENTRAL DET
ARTICLE
the geneticist L.L. Cavalli-Sforza , and the anthropologist F.T. Clark and the ethologist J.M. Cullen

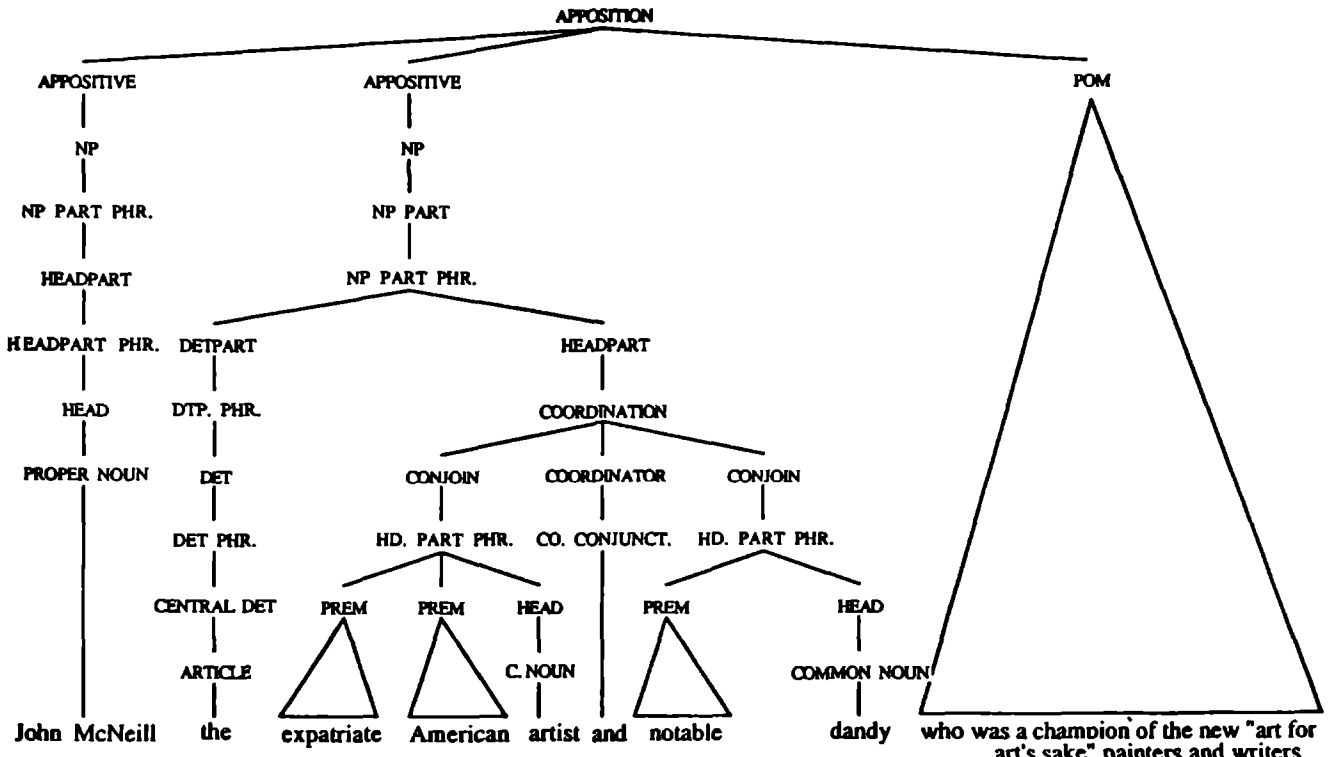

APPOSITION
APPOSITIVE
APPOSITIVE
POM
NP
NP
NP PART PHR.
NP PART
HEADPART
NP PART PHR.
HEADPART PHR.
DETPART
HEADPART
HEAD
DTP. PHR.
COORDINATION
PROPER NOUN
DET
CONJOIN
COORDINATOR
CONJOIN
DET PHR.
HD. PART PHR.
CO. CONJUNCT.
HD. PART PHR.
CENTRAL DET
PREM
PREM
HEAD
PREM
HEAD
ARTICLE
C. NOUN
COMMON NOUN
John McNeill
the
expatriate
American
artist
and
notable
dandy
who was a champion of the new "art for art's sake" painters and writers

Figure 38

4. focusing adjuncts do not co-occur. Other adjuncts can, however, be the focus of restrictive adverbials such as *only*; cf.

(113) They *only* want the car *for an hour*.
* (114) They *only* want the car *even*.

5. focusing adjuncts that constitute single adverbs cannot be premodified by *however, how*, or *so* whereas other adverb adjuncts can. For example,

(115) However *strongly* you feel about it, you should be careful what you say.
(116) How *often* does he drink beer?
(117) How *cautiously* he drives!
(118) So *monotonously* did he speak that everyone left.

The fact that focusing adjuncts behave differently, syntactically speaking, from other adjuncts, need not necessarily lead to the conclusion that therefore they cannot be dealt with as adverbials on clause or sentence level, although in numerous instances this results in rather awkward analyses. Much more conclusive in deciding that an alternative description for focusing adjuncts is desired is the fact that they constitute a fairly limited set of items that can be found to occur in coordinations and appositions in positions where other adjuncts do not normally occur. Consider for example,

(119) *Peter and even Bob* was there.
(120) *These workers, chiefly dockworkers,* are dissatisfied with the government.

Two possibilities must be considered for describing instances like these. The first, describing focusing adjuncts as part of the coordination or apposition, is considered rather unattractive because this would result in an inconsistency between the analysis of single noun phrases (i.e. noun phrases without coordination or apposition) on the one hand, and noun phrases in which coordination or apposition is found to occur on the other hand. The other possibility, then, is to describe focusing adjuncts as part of the noun phrase. This makes it possible to look upon *only the extremely wealthy customers* (99), *at least ten workers* (100), *especially the girls* (101), *the workers in particular* (102), *even Bob* (103), as single constituents (noun phrases), while *Peter and even Bob* (119), and *these workers, chiefly dockworkers* (120) can be looked

upon as a coordination of noun phrases or an apposition of noun phrases respectively. Thus the analysis is consistent, irrespective of the composition of the noun phrase. Of course in this approach the analysis of instances like (104), in which the focusing adjunct is not juxtaposed to the noun phrase on which it focuses, remains problematic. It is for these cases that we propose the analysis of the focusing adjunct as adverbial on sentence or clause level.

Floating postmodifiers

It was observed earlier that the description of focusing adjuncts that are not juxtaposed to the noun phrase on which they focus is problematic. The same applies to noun phrase postmodifiers that are found to be postponed. Consider the following examples (taken from Quirk et al., 1972: 966):

(121) *The time* had come *to decorate the house for Christmas.*
(122) *That loaf* was stale *that you sold me.*
(123) *What business* is it *of yours*?
(124) *All of us* were frightened *except the captain.*

Note that on the whole noun phrase postmodifiers are problematic, especially when occurring in sentence or clause final positions, in that it appears very difficult indeed, if not impossible, to distinguish (on syntactic grounds) between postmodifiers and adverbials.

Deferred determiners

Given the fact that some determiners are found to occur immediately following the head they qualify, our description of the noun phrase must be adapted so as to accommodate these deferred determiners. However, we also come across other variants where the determiners are deferred beyond the position immediately following the head. Consider the following examples:

(125) *All the authors* were invited.
(126) *They all* were invited.
(127) *The authors* were *all* invited.

Items that can occur as deferred determiners are few: only the universal pronouns *all, each* and *both* can do so.

4.4 Concluding remarks

In this chapter some aspects of the implementation of the grammar have been discussed. Since it would have gone too far to describe each and every detail of the grammar as it is available today, we have merely tried to highlight some of the issues that we came across while writing the grammar. It should be noted, therefore, that in this chapter we have had no intention of being exhaustive in our description. Rather, the discussion of coordination and gapping, and the noun phrase should give the reader an idea of how a formalized description of the syntax of a language may be arrived at and what considerations may play a role. Adapted versions of the rules presented in this chapter occur in today's grammar as it is used in the analysis of the TOSCA corpus. Between the first conception of the rules and their ultimate incorporation in the grammar lay a period of quite some time, during which modules of the grammar were written and tested, revised and integrated into other modules. An evaluation of the grammar is given in the next chapter, together with an overview of some residual problems that are yet to be dealt with.

5.0 Evaluation and Conclusion

5.1 Introductory

In this final chapter an evaluation is given of the functioning of the grammar as it was employed in the analysis of the corpus of English described in chapter 2. We include a discussion of some of the analysis results obtained and of residual problems to be dealt with. While the assessment of the grammar and its performance that is given in section 5.3 remains informal, section 5.4 focuses on the question of how a more formal standard could be set for assessing the grammar. We conclude this chapter with a brief discussion of what still has to be done, indicating the direction future research should take.

An (informal) evaluation of the grammar can only properly be made by taking into consideration various factors that have played a role in the way the grammar was set up. Among these factors are the choices that guided the design of the grammar and its further development, but also circumstances that simply existed and could not (easily) be altered. To be more precise, the grammar that was written for the analysis of the TOSCA Corpus must be evaluated in the light of

1. the priorities that were set (cf. chapter 3); thus the creation of a database in which all grammatical strings had been analyzed was given priority over other aims;

2. the fact that, for the time being, it was preferred to keep close to what was traditional and familiar in linguistic description; one of the primary goals in yielding an analyzed corpus was to make available a store of data that would be easily accessible for a great many linguists from various backgrounds; it was, therefore, considered to be of some importance to try and have analyses that could be readily interpreted;

3. the restrictions that were adopted; the scope of the grammar was restricted to the morpho-syntactic analysis of the individual utterances in a text;

4. the developments taking place in the fields of computer science and modern (corpus) linguistics; here it must be observed that far from being stable, the research environment was constantly developing; in the same way that the corpus linguists had to grow to their tasks (writing the grammar, incorporating newly acquired insights, etc.), the application of approved computer scientific methods and techniques in the field of corpus linguistics continued to pose a challenge to the computer scientists as complexes of problems occurred and solutions had to be worked out to enable the analysis to proceed within acceptable boundaries.

Moreover, it ought to be kept in mind that the grammar was developed in the course of a research project that aimed at the detailed syntactic analysis of a one million word-corpus; given the primary aim of the project it is clear that at a certain point in time one cannot afford to continue adapting the grammar, since the analysis of the corpus can no longer be postponed; whatever insights are gained in the process must await future revision and adaptation of the grammar and the linguistic hypotheses incorporated in it.

All in all, as developments occur, expertise increases, and we are capable of extending our goals, it is only to be expected that in retrospect some things could have been dealt with more adequately, if only we had had the means and expertise that are available to us now. Also, we should realize that even at present our speed of operation, working, as we do, with unrestricted input, does not allow for any circularity in the writing and testing of a grammar of this size and nature on a corpus like the TOSCA Corpus.

5.2 Intermezzo: some analysis results

On the whole the performance of the grammar is quite satisfactory. In order to give some idea of the nature of the material and of the analyses and the amount of detail in them, we include some excerpts from one of the corpus samples[1] together with the analyses that were obtained.

[1] The sample in question falls within the text category 'crime fiction' and was taken from Michael Innes' *Carson's Conspiracy. A Sir John Appleby Mystery.* The excerpts that are included here all occur in chapter 9. The exact references are: excerpt 1, page 91; excerpt 2, page 95; and excerpt 3, page 102. The location codes that occur at the beginning of each line in

excerpt 1

*<*29*>
|
|F*0or some days, however, nothing of the sort happened. Somewhat sporadically at this time, Appleby was writing a book. It wasn't autobiographical, and such sensational crimes as it touched on had occurred for the most part in the fourteenth and fifteenth centuries. Appleby had taken to that investigating and recording of local history which has become prominent as an unassuming pursuit among the elderly and literate classes. When questioned about it, he would say that it served as well as the bees. This was understood to be an allusion to the final phase in the career of Sherlock Holmes.

The representation of the original text as it was keyed onto tape required decisions about what to retain and what to leave out. Material that was not included was either represented by tags (this was the case with such matter as figures, diagrams, tables, quotations, mathematical formulae, etc.) or was excluded without further reference (this applied to notes, annotations that occurred in margins, headers and footers, references, etc.). Also we decided to have some sort of coding of (mainly) the typographical features of the text. Thus changes in font type were encoded, as were paragraph and dialogue indentation, blank spaces or lines, headings, abbreviations, foreign material, etc.[2]

When it came to analyzing the material a distinction was made between textual units, headings, markup and extratextual material. Contrary to markup and extratextual material which were to receive only a trivial analysis on the basis of the grammar, textual units and --

---

the computer readable version of the text are not given here.

[2] The coding key for this material may be found in the *TOSCA Corpus -- Manual* (Oostdijk, 1989). Here we restrict ourselves to presenting a key to just the codes that occur in the excerpts included here.

*<	begin heading	*-	dash
*>	end heading	*"	opening double quote
*0	begin roman	**"	end double quote
*1	begin italics	*&	begin dialogue indentation
*2	begin bold face type	|	blank line(s); begin new paragraph
*3	begin bold face italics		

to some extent -- headings were subject to a detailed syntactic analysis.

The analysis process may be divided in the following three steps: (1) tokenization, (2) lexical-morphological analysis, and (3) syntactic analysis. Each of these is discussed below.

The first step in the analysis process, tokenization, consists in separating the various units that are distinguished. Thus paragraph markers, dialogue indentation markers, etc. are separated from the text they precede, and individual textual units are identified. As a result of this first step in the tokenization process the text found in excerpt 1 looks as follows:

*<*29*>
¬
|
¬
|
¬
F*0or some days, however, nothing of the sort happened.
¬
Somewhat sporadically at this time, Appleby was writing a book.
¬
It wasn't autobiographical, and such sensational crimes as it touched on had occurred for the most part in the fourteenth and fifteenth centuries.
¬
Appleby had taken to that investigating and recording of local history which has become prominent as an unassuming pursuit among the elderly and literate classes.
¬
When questioned about it, he would say that it served as well as the bees.
¬
This was understood to be an allusion to the final phase in the career of Sherlock Holmes.

A second step in the tokenization process constitutes the identification and separation of the indivual tokens. For example,

*<*29*>

is converted to

```
*<
*29
*>
```

and

F*0or some days, however, nothing of the sort happened.

is converted to

```
F*0or
some
days
,
however
,
nothing
of
the
sort
happened
.
```

The tokenization process is fairly straightforward and presents only a few minor problems. For instance, the output of the tokenizer is ambiguous in some cases, and occasionally tokenization fails. As far as the identification of textual units is concerned, an ambiguous result is obtained for instances, especially in reported utterances, where one of the following sequences of characters is found[3]

... followed by a blank and a capital letter
?' followed by a blank and a capital letter
?" followed by a blank and a capital letter
!' followed by a blank and a capital letter
!" followed by a blank and a capital letter

For example, given texts (a) and (b) as input, the tokenizer produces a similar (ambiguous) result for each, while also the tokenization of text (c) yields an ambiguous result.

[3] N.B. This list is not intended to be exhaustive.

(a) "Do you know who they are?" Appleby asked.
(b) "Mr. and Mrs. Lely, you mean?" Appleby wondered whether Hoobin ought to be reproved for calling Humphry Lely an artist creature, but decided that nothing markedly derogatory had been intended.
(c) "Have you any evidence for that?" It seemed remarkable to Appleby that there should be the same presage of improbably drastic doom at Garford House as he had received little more than an hour ago from William Lockett himself.

The identification of the individual tokens is unambiguous with the one notable exception of line-final hyphenated tokens. The tokenizer cannot distinguish between words that are always hyphenated (irrespective of the context they occur in), and words that are not normally hyphenated, but are simply broken off at the end of a line. Ambiguity in this phase of the analysis typically results from the fact that the tokenization precedes any dictionary lookup.

During the next step in the analysis process, the lexical-morphological analysis, each token is tagged on the basis of a computer readable dictionary that has been adapted for the purpose and includes a morphological component. The total number of distinct wordtypes amounts to about 70,000.[4] Lexical items are assigned tags indicating their word class membership and such properties as number, complementation type, tense, etc. A separate set of tags exists for the tagging of punctuation marks, extratextual material and markup. Punctuation marks are tagged PUNCM or IGN ('ignore'). PUNCM is assigned if the punctuation mark plays a role on the syntactic level, IGN if it has no such function. For example, a comma in a text is tagged PUNCM if it functions as a coordinator or as an end-of-(reported) utterance marker. If, on the other hand, a comma indicates a pause, it is tagged IGN.

The result of the lexical-morphological analysis for our example looks as follows:[5]

---

[4] Of the 55,000 entries (headwords and derivatives) that occurred in the original version of the computer-readable dictionary any multiple entries for the same wordtype were conflated so that for each wordtype only one entry remained. The application of a set of morphological rules yielded a wordlist consisting of some 70,000 distinct wordtypes. Each entry was provided with wordclass and feature information.

[5] The following key applies to the tags that were used in the examples given. Here capitalized

```
*<          MUP(ohead)
*29         CARD(plu)
*>          MUP(chead)

F*0or       PREP;COCO
some        DET(ass,cnty);PN(ass,number)
days        CN(plu)
,           IGN;PUNCM(com)
however     CON
,           IGN;PUNCM(com)
nothing     PN(neg,sing)
of          PREP
the         ART(cnty);ADV(in,abs)
sort        CN(sing);MLV(motr,infin);MLV(motr,pres)
happened    MLV(intr,past);MLV(intr,pastp)
            PUNCM(per)
```

abbreviations indicate word class categories, while features are between brackets (using small letters).

Categories:

ART	article	IGN	ignore
ADV	adverb	MLV	main lexical verb
CARD	cardinal numeral	MUP	markup
CN	common noun	PN	pronoun
COCO	coord. conjunction	PREP	preposition
CON	connective	PUNCM	punctuation mark
DET	determiner		

Features:

ass	assertive	neg	negative
abs	absolute	number	number
chead	close head	ohead	open head
cnty	countability	past	past
com	comma	pastp	past participle
in	intensifying	per	period
infin	infinitive	plu	plural
intr	intransitive	prep	prepositional
modal	modal	pres	present
motr	monotransitive	sing	singular

While generally in the tokenization process there is no need to intervene since further steps in the analysis restrict the ambiguity, interventions are needed to resolve part of the ambiguity which results from the lexical-morphological tagging before further analyzing a string. This *lexical distributional ambiguity*, i.e. the ambiguity with respect to the word class and/or feature set of a particular token, can (at least partly) be solved by the syntactic parser. However, we have found that although the parser may be capable of such disambiguation, leaving the parser to sort things out by itself is extremely costly both in terms of computer time and memory space. Here it must be observed that, in general, the disambiguation of the lexical-morphological tags through intervention is much more powerful than any disambiguation achieved by the syntactic parser on this point. Whereas the parser succeeds in resolving only part of the ambiguity, the disambiguation through intervention is total. For example, the lexical-morphological tagging of the string 'They can fish' yields the following result:

They	PN(per,plu)
can	AUX(modal,pres);MLV(motr,infin);MLV(motr,pres);CN(sing)
fish	CN(sing);CN(plu);CN(mass);MLV(intr,infin);MLV(intr,pres); MLV(motr,infin);MLV(motr,pres)

On the basis of this input the syntactic parser arrives at a fourfold ambiguous analysis for this string. While it discards the tags CN(sing) and MLV(motr,infin) for *can* and also MLV(intr,pres), MLV(motr,infin) and MLV(motr,pres) for *fish*, the tagging for both *can* and *fish* remains ambiguous. Co-occurrence restrictions that are expressed in the syntactic rules prohibit the analysis of both *can* and *fish* as MLV, so that the following analyses remain:

(a) PN(per,plu) - AUX(modal,pres) - MLV(intr,infin)
(b) PN(per,plu) - MLV(motr,pres) - CN(sing)
(c) PN(per,plu) - MLV(motr,pres) - CN(plu)
(d) PN(per,plu) - MLV(motr,pres) - CN(mass)

If we were to disambiguate the tagging through intervention, by selecting only the contextually appropriate tags, this would result in a single tag for each token. In the case of our example the syntactic analysis then is no longer ambiguous. Note that in other instances the result of the syntactic parser (when provided with unambiguous, i.e. fully disambiguated, strings as input) may still be syntactically ambiguous, even when any residual lexical distributional ambiguity has been removed.

Since any lexical distributional ambiguity that the syntactic parser fails to resolve may give rise to further syntactic ambiguity, the disambiguation of the tagging through intervention (i.e. selection) indirectly causes the syntactic analysis to be less ambiguous.

Apart from the lexical distributional ambiguity that arises from the lexical-morphological analysis, there is a second type of ambiguity which is typically generated by the syntactic parser and therefore referred to as *syntactic ambiguity*. Syntactic ambiguity arises when lexical items whose word class membership is unambiguous can be grouped together in different ways. In the case of our (lexically disambiguated) example string the following bracketings are found:

((F*0or some days,)(however,)(nothing of the sort)(happened)(.))
((F*0or some days,)(however,)(nothing)(of the sort)(happened)(.))

Here the amount of ambiguity generated by the syntactic parser is manageable. Yet with more complex input syntactic ambiguity appears not only the most frequent but also the most troublesome kind of ambiguity that we come across in the process of analyzing a corpus. For instance, without further semantic knowledge it appears impossible to distinguish between modifying and adverbial constituents, i.e. the recognition of a word group as either IC of a sentence or clause, or IC of a phrase (hence the ambiguity in the example above). Given the present limitations of corpus linguistic practice -- its abstraction from context and situation, and its restriction to a syntactic analysis -- there is little else the linguist can do but intervene. At this point it becomes apparent that intervention is necessary. Even with strings of disambiguated tags parsing times may sometimes get completely out of hand, while occasionally it also happens that disk space runs out. At the present time we have therefore opted for resolving some problems of attachment by applying a semantically and pragmatically based pre-analysis, i.e. indicating the boundaries of certain constituents. In practice this means that for the utterances in excerpt 1 a pre-analysis carried out by the linguist yields the following constituent boundary marking[6] (here indicated by means of square brackets):

[6] The syntactic markers that were used are discussed extensively in Appendix H. Here it should be observed that we have opted for a minimal marking so that a constituent boundary marking generally consists of a single bracket marking either the beginning or the end of a constituent, whichever is required. Only with some constituents bracket pairs are used.

F*0or some days, however, nothing of the sort] happened.

Somewhat sporadically at this time, Appleby was writing a book.

It wasn't autobiographical, and such sensational crimes [as it touched on] had occurred for the most part in the fourteenth and fifteenth centuries.

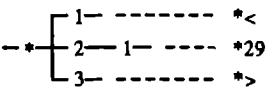
Appleby had taken to that investigating and recording of local history] which has become prominent as an unassuming pursuit among the elderly and literate classes].

When questioned about it, he would say that it served as well as the bees.

This was understood [to be an allusion to the final phase in the career of Sherlock Holmes]]].

For the text in excerpt 1, the analyses as they were obtained in the analysis process described above and subsequently stored in the Linguistic Database (LDB) are given below. The analyses are represented in the form of tree diagrams. Unlike the tree diagrams usually found in linguistic studies, these trees grow from left to right. The leftmost node constitutes the root of the tree, while on the rightmost nodes the lexical categories can be found. Trees in the LDB can be viewed in two different modes: the tree map view and the environment view. The tree map view shows the overall structure of an analysis, while the environment view presents a more detailed representation of the information that is available for each of the nodes in a tree. In the examples given below the tree map view of each utterance is presented in LDB format. The information contained in the environment view, however, is not given in LDB format, since analysis trees tend to become rather large when viewed in this mode. Instead, this information is included in the indented representations that are given below the tree structures. The information is structured as follows: function-category pairs occupy a single line and may be followed by information of a syntactico-semantic nature which is enclosed between brackets; lexical elements occur between braces. A key to the function and category labels, and also the features that are associated with these, can be found in Appendix F.

Figure 1: *<*29*>

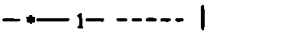



```
NOFU,HEAD()
: HDIN,HDMO(OHEAD){*<}
: SPEC,TXTU()
:   UTT,NP()
:     NPHD,NN(){*29}
: HDTL,HDMO(CHEAD){*>}
```

The analysis of this heading is fairly trivial since it consists only of a single item. The noun phrase is one of the more frequent realizations of the heading, and although often fairly simple in structure such noun phrases occasionally become quite complex; for example,

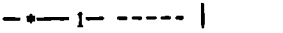
(1) Tests for discrimination -- conditioned reflexes
(2) Chapter 11 Selection 3: The Study of Learners' Language: Error Analysis

Figure 2: |




```
NOFU,MUP(BLANK){|}
```

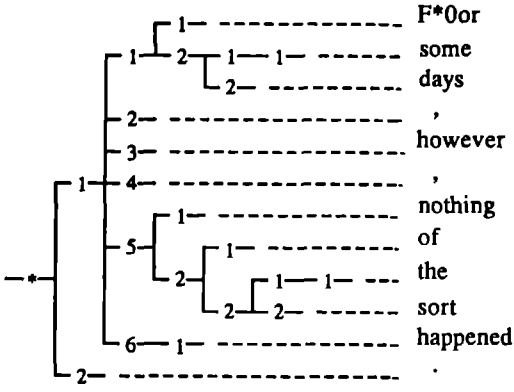
Figure 3: |




```
NOFU,MUP(PAR){|}
```

The analysis of the markup as found in Figures 2 and 3 is straightforward. The two different uses of the vertical bar (|) as coding symbol are exemplified here: in Figure 2 it represents one or more blank lines in the original text, while in Figure 3 it indicates the beginning of a paragraph.

Figure 4: F*0or some days, however, nothing of the sort happened.




```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   A,PP()
:     P,PREP()(F*0or)
:     PC,NP()
:       :DT,DTP(PLU)
:       :   DTCE,DET(ASS,PLU)(some)
:       :NPHD,CN(PLU)(days)
:   NOFU,NOCA(IGN)(,)
:   ADCO,CON()(however)
:   NOFU,NOCA(IGN)(,)
:   SU,NP()
:     NPHD,PN(NEG,SING)(nothing)
:     NPPO,PP()
:       :P,PREP()(of)
:       :PC,NP()
:       :   DT,DTP(SING)
:       :     DTCE,ART(SING)(the)
:       :   NPHD,CN(SING)(sort)
:   VB,VP(INTR)
:     MVB,MLV(INTR,PAST)(happened)
: PUNC,PUNCM(PER)(.)
```

The analysis represented in Figure 4 is that of a textual unit (TXTU) which consists of an utterance (UTT) followed by a punctuation mark (PUNC). The utterance is realized by a regular declarative sentence 'S(REG,DECL)'. Within this sentence four function slots are distinguished; the commas that occur in this sentence are ignored in the syntactic analysis and are assigned the dummy labels NOFU ('no function') and NOCA ('no category'). The sentence pattern found in this sentence is that of an adverbial (A), followed by a connective adjunct (ADCO), followed by a subject (SU) and a verb (VB).

The adverbial is realized by a prepositional phrase (PP) which consists of a preposition (P) followed by a prepositional complement (PC). The function preposition is realized by a preposition (PREP), which in this case is *F*0or.*[7] The prepositional complement is realized by an NP, which consists of a determiner (DT) followed by a head (NPHD). With the determiner phrase (DTP) which realizes the function of determiner, a feature 'plural' (PLU) is associated. The central determiner is the only function in the determiner phrase; it is realized by the assertive determiner *some.*

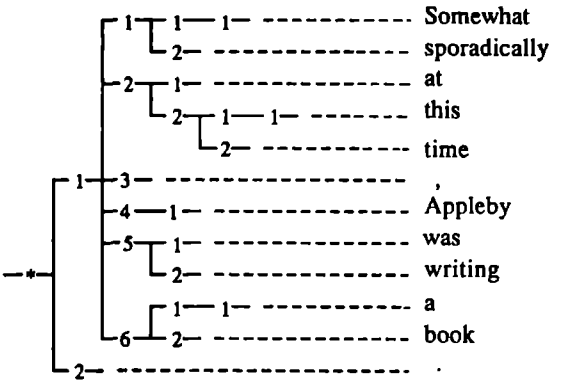
The connective adjunct is realized by a connective (CON). Apart from typical connective items such as *however, moreover, therefore, first(ly),* and *on the one/other hand,* the class of connectives also includes items like *and, or, but, not* and *neither,* whenever these occur sentence-initially.

The subject is realized by an NP, which consists of a head followed by a postmodifier (NPPO). The head is realized by the negative pronoun *nothing.* The postmodifier is realized by a prepositional phrase. The preposition *of* is followed by an NP (*the sort*).

Finally, the verb is realized by an intransitive verb phrase (VP) The sole constituent here is the main verb (MVB), realized by the main lexical verb (MLV) *happened.*

[7] Although information about changes in font type was retained by encoding them (by means of an asterisk code, here *0), this is ignored in the analysis.

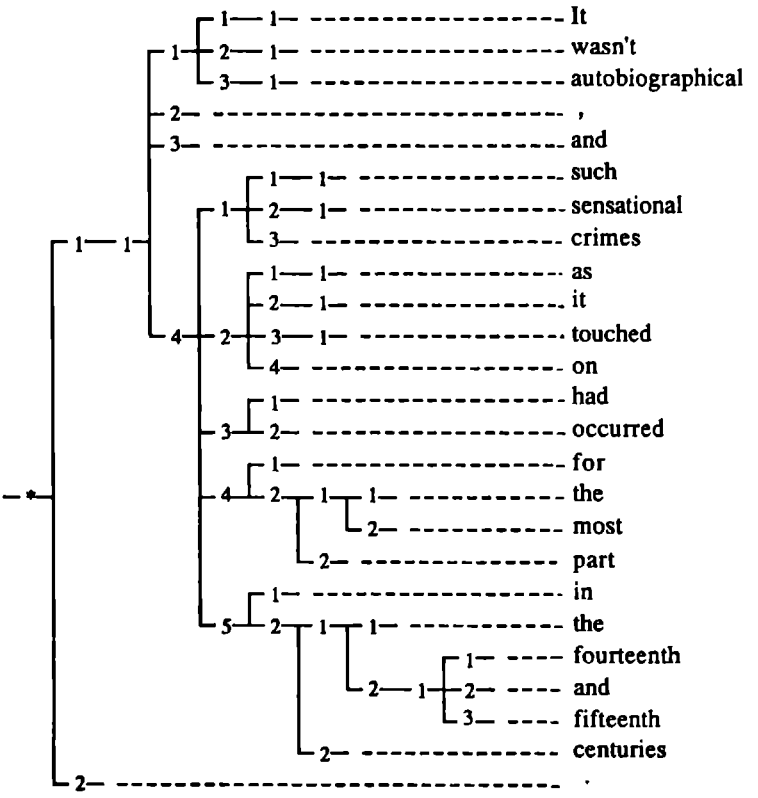
Figure 5: Somewhat sporadically at this time, Appleby etc.




```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   A,AVP(GE)
:     AVPR,AVP(GE)
:       :AVHD,ADV(GE,ABS){Somewhat}
:     AVHD,ADV(GE,ABS){sporadically}
:   A,PP()
:     P,PREP(){at}
:     PC,NP()
:       :DT,DTP(SING)
:       :   DTCE,DET(DEM,SING){this}
:       :NPHD,CN(SING){time}
:   NOFU,NOCA(IGN){,}
:   SU,NP()
:     NPHD,PRN(SING){Appleby}
:   VB,VP(MOTR)
:     AVB,AUX(PROG,PAST){was}
:     MVB,MLV(MOTR,PRESP){writing}
:   OD,NP()
:     DT,DTP(SING)
:       :DTCE,ART(SING){a}
:     NPHD,CN(SING){book}
: PUNC,PUNCM(PER){.}
```

The analysis represented in Figure 5 shows two adverbials, one realized by an adverb phrase (AVP) and the other realized by a prepositional phrase (PP). Note that within the adverb phrase another adverb phrase is found as premodifier (AVPR).

Figure 6: It wasn't autobiographical, and such etc.




```
NOFU,TXTU()
: UTT,NOCA()
:   NOFU,NOCA(COORD)
:     CJ,S(REG,DECL)
:       :SU,NP()
```

```
:       :   NPHD,PN(PERS,SING){It}
:       :VB,VP(INTENS)
:       :   MVB,MLV(INTENS,NEG,PAST){wasn't}
:       :CS,AJP()
:       :   AJHD,ADJ(ABS){autobiographical}
:      NOFU,NOCA(IGN){,}
:      COOR,COCO(){and}
:      CJ,S(REG,DECL)
:       :SU,NP()
:       :   DT,DTP(PLU)
:       :      DTPS,QUANT(PLU){such}
:       :   NPPR,AJP()
:       :      AJHD,ADJ(ABS){sensational}
:       :   NPHD,CN(PLU){crimes}
:       :FDTPO,FC(SUBORD)
:       :   SUB,SUBP()
:       :      SUBHD,COSU(){as}
:       :   SU,NP()
:       :      NPHD,PN(PERS,SING){it}
:       :   VB,VP(INTR)
:       :      MVB,MLV(INTR,PAST){touched}
:       :   A,AVP(PREP)
:       :      AVHD,ADV(PREP){on}
:       :VB,VP(INTR)
:       :   AVB,AUX(PERF,PAST){had}
:       :   MVB,MLV(INTR,PASTP){occurred}
:       :A,PP()
:       :   P,PREP(){for}
:       :   PC,NP()
:       :      DT,DTP(SING)
:       :         DTCE,ART(SING){the}
:       :         DTPS,DET(ASS,SING){most}
:       :      NPHD,CN(SING){part}
:       :A,PP()
:       :   P,PREP(){in}
:       :   PC,NP()
:       :      DT,DTP(PLU)
:       :         DTCE,ART(PLU){the}
:       :         DTPS,NOCA()
:       :            : NOFU,NOCA(COORD)
:       :            :    CJ,ORD(PLU){fourteenth}
```

```
:       :       :    COOR,COCO()(and)
:       :       :    CJ,ORD(PLU)(fifteenth)
:       :     NPHD,CN(PLU)(centuries)
: PUNC,PUNCM(PER)(.)
```

The utterance in Figure 6 is realized by a coordination of two sentences. Since the realization of the function (in this case the function 'utterance') does not constitute a single constituent, but is a coordination of multiple categories, the 'categorial' label that is associated with this node is that of 'NOCA' (no category).[8] The labelling 'NOFU, NOCA(COORD)' associates the feature 'coordination' (COORD) with the intermediate node, which is introduced to indicate the fact that a coordination occurs here. Note that no function or category is associated with this intermediate node. The functions that are distinguished within the coordination are conjoin (CJ) and coordinator (COOR). Both conjoins are realized by a regular declarative sentence. The sentence pattern of the first is that of a subject followed by a verb followed by a subject complement (CS). In the second sentence, we have a subject followed by a verb followed by two adverbials. However, the determiner phrase of the subject NP appears to be discontinuous: while *such* precedes the premodifier and the head, the continuation of this determiner phrase follows the head. Since we do not attempt, in our description, to associate such floating constituents with their parents (other than in the labelling), in the analysis they occur as independent constituents. The floating determiner phrase postmodifier (FDTPO) found in this analysis is a typical example. The decision to analyze floating constituents as independent constituents is motivated by the fact that the 'mobility' of these constituents does not appear to be restricted to the next highest level. For instance, the floating determiner phrase postmodifier does not necessarily occur within the boundaries of the NP (which is the next highest constituent):

(3) But so many people have, *that the fact is not notable in itself.*

In example (3) the floating determiner phrase postmodifier does not occur as an immediate constituent of the NP. Instead, it occurs as an IC of the sentence.

[8] The realization of a function by multiple categories occurs with instances of coordination (as here) and apposition.

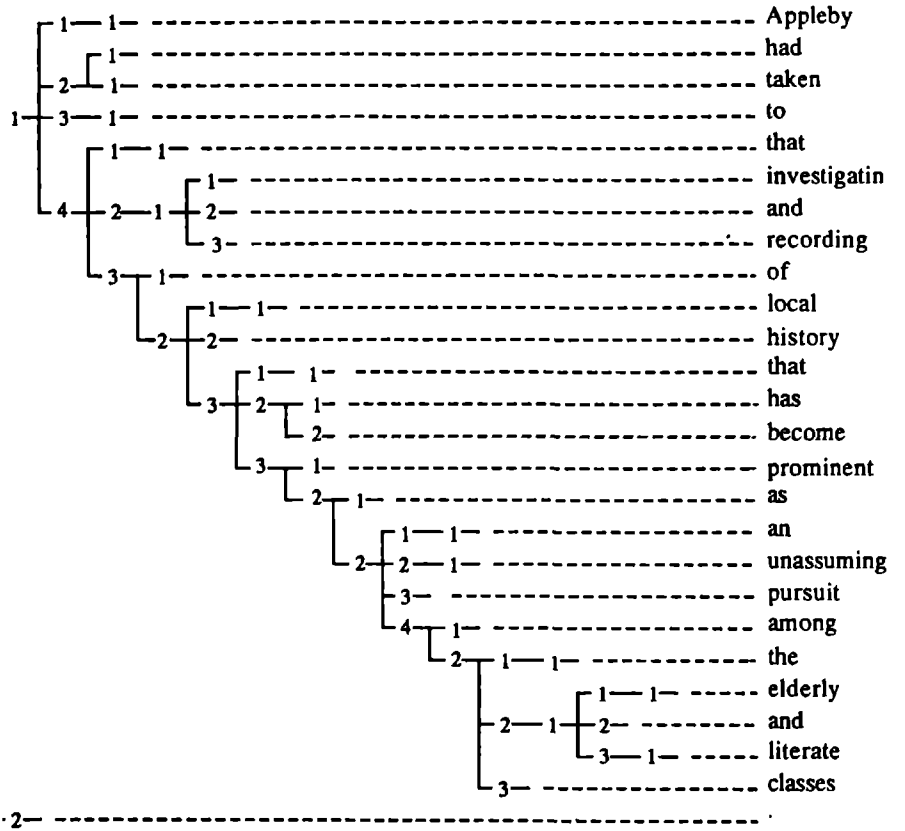
The analysis of prepositional verbs and phrasal verbs poses a problem. When we attempt to design a description that assigns the label 'prepositional verb' to a combination of a verb and a preposition that goes with this verb, or the label 'phrasal verb' to a combination of a verb and an adverb, it appears that such a description tends to become extremely complicated in those instances where the preposition or the adverb does not immediately follow the verb it belongs to.[9] Since we want our grammar to yield the same analysis in all instances, i.e. irrespective of whether the preposition or the adverb does or does not immediately follow the verb, we opted for a description which would distinguish between the verbal part on the one hand and the preposition or the adverb on the other hand. The analysis of phrasal verbs in these terms is unproblematic. With prepositional verbs, however, this implies that we admit prepositional phrases that consist of a single constituent: the preposition. The prepositional complement in this case would become an optional constituent. As a side-effect of abandoning the description of the prepositional phrase as an exocentric construction the analysis of prepositional phrases would become ambiguous. Therefore the following solution was adopted. The description of prepositional phrases remained what it was; with prepositional verbs the verbal part is analyzed as verb and the preposition receives the label 'prepositional adverb' (ADV with the feature 'PREP'). In (7) an example of this kind of analysis is found for *had taken to*.

[9] Consider the examples presented by Quirk et al. (1972: 816):

(a) They call early on the man
(b) They call him up.

In the first example we find the prepositional verb 'call on', while example (b) contains the phrasal verb 'call up'. Both verbs are monotransitive; 'the man' (a) and 'him' (b) are direct objects.

igure 7: Appleby had taken to that investigating etc.

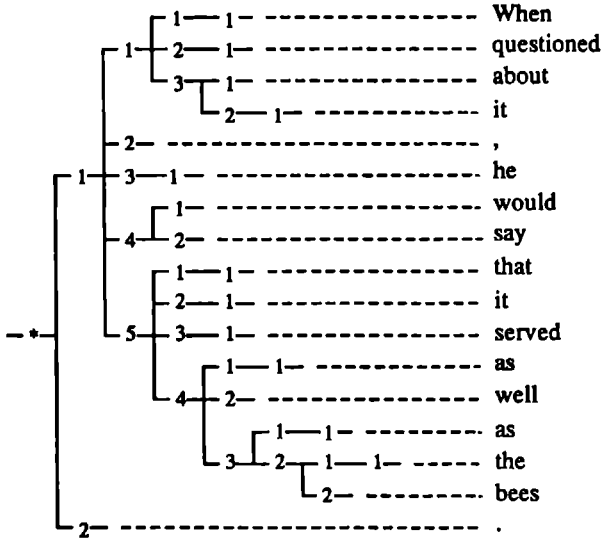



```
OFU,TXTU()
 UTT,S(REG,DECL)
   SU,NP()
     NPHD,PRN(SING){Appleby}
   VB,VP(MOTR)
     AVB,AUX(PERF,PAST){had}
     MVB,MLV(MOTR,PASTP){taken}
   A,AVP(PREP)
     AVHD,ADV(PREP){to}
```

```
:  OD,NP()
:    DT,DTP(MASS)
:      :DTCE,DET(DEM,MASS){that}
:    NPHD,NOCA()
:      :NOFU,NOCA(COORD)
:      :  CJ,NPHD(MASS){investigating}
:      :  COOR,COCO(){and}
:      :  CJ,NPHD(MASS){recording}
:    NPPO,PP()
:      :P,PREP(){of}
:      :PC,NP()
:      :  NPPR,AJP()
:      :    AJHD,ADJ(ABS){local}
:      :  NPHD,CN(MASS){history}
:      :  NPPO,FC(REL)
:      :    SUB,SUBP()
:      :      SUBHD,PN(REL,SING){that}
:      :    VB,VP(INTR)
:      :      AVB,AUX(PERF,PRES){has}
:      :      MVB,MLV(INTENS,PASTP){become}
:      :    CS,AJP()
:      :      AJHD,ADJ(ABS){prominent}
:      :      AJPO,PP()
:      :      : P,PREP(){as}
:      :      : PC,NP()
:      :      :   DT,DTP(SING)
:      :      :     DTCE,ART(SING){an}
:      :      :   NPPR,AJP()
:      :      :     AJHD,ADJ(ABS){unassuming}
:      :      :   NPHD,CN(SING){pursuit}
:      :      :   NPPO,PP()
:      :      :     P,PREP(){among}
:      :      :     PC,NP()
:      :      :       :DT,DTP(PLU)
:      :      :       :  DTCE,ART(PLU){the}
:      :      :       :NPPR,NOCA()
:      :      :       :  NOFU,NOCA(COORD)
:      :      :       :    CJ,AJP()
:      :      :       :      AJHD,ADJ(ABS){elderly}
:      :      :       :    COOR,COCO(){and}
:      :      :       :    CJ,AJP()
```

```
:       :       :       :       AJHD,ADJ(ABS){literate}
:       :       :       :NPHD,CN(PLU){classes}
: PUNC,PUNCM(PER){.}
```

Figure 8: When questioned about it, he would say etc.




```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   A,NFC()
:     SUB,SUBP()
:       SUBHD,COSU(){When}
:     VB,VP(INTR)
:       :MVB,MLV(INTR,PASTP){questioned}
:     A,PP()
:       :P,PREP(){about}
:       :PC,NP()
:       :   NPHD,PN(PERS,SING){it}
:   NOFU,NOCA(IGN){,}
:   SU,NP()
:     NPHD,PN(PERS,SING){he}
:   VB,VP(MOTR)
```

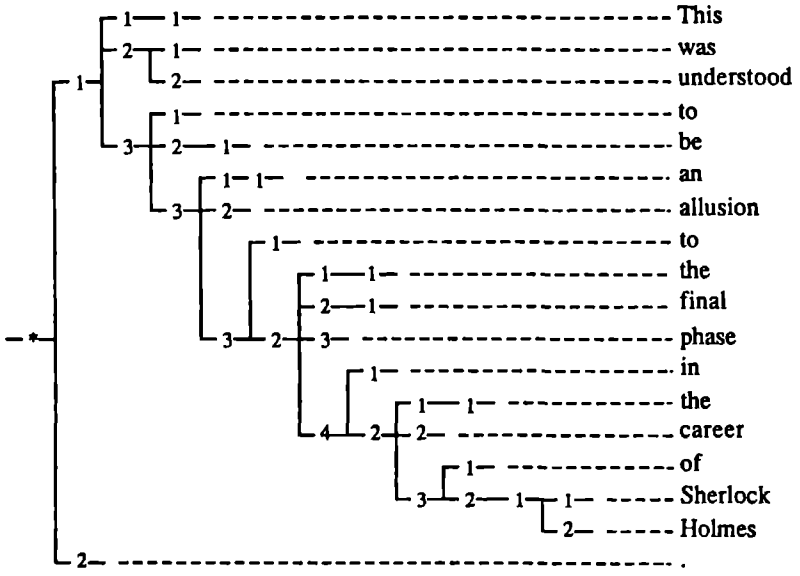
```
:      AVB,AUX(MODAL,PAST){would}
:      MVB,MLV(MOTR,INFIN){say}
:   OD,FC(SUBORD)
:      SUB,SUBP()
:       :SUBHD,COSU(){that}
:      SU,NP()
:       :NPHD,PN(PERS,SING){it}
:      VB,VP(INTR)
:       :MVB,MLV(INTR,PAST){served}
:      A,AVP(GE)
:       :AVPR,AVP(IN)
:       :  AVHD,ADV(IN){as}
:       :AVHD,ADV(GE){well}
:       :AVPO,FC(RED)
:       :  SUB,SUBP()
:       :    SUBHD,COSU(){as}
:       :  SU,NP()
:       :    DT,DTP(PLU)
:       :      :DTCE,ART(PLU){the}
:       :    NPHD,CN(PLU){bees}
: PUNC,PUNCM(PER){.}
```

The analysis represented in Figure 8 exemplifies the analysis of what we have termed a 'reduced finite clause' (FC with the affix 'RED'). A reduced finite clause is a clause that is introduced by a subordinator; the subject is the only other obligatory constituent. This type of clause is called a reduced *finite* clause rather than a (reduced) non-finite clause or verbless clause since it can only be extended by means of a finite verb phrase. Reduced finite clauses typically occur in comparative constructions.[10] For example,

(4) He ran faster *than any of his friends.*
(5) They can do this as easily *as anyone else.*

[10] Aarts and Aarts (1982) refer to these clauses as '(reduced) comparative clauses'.

Figure 9: This was understood to be an allusion etc.




```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   SU,NP()
:     NPHD,PN(DEM,SING){This}
:   VB,VP(INTENS)
:     AVB,AUX(PASS,PAST){was}
:     MVB,MLV(CXTR,PASTP){understood}
:   CS,NFC()
:     PART,PRTCL(TO){to}
:     VB,VP(INTENS)
:      :MVB,MLV(INTENS,INFIN){be}
:     CS,NP()
:      :DT,DTP(SING)
:      :  DTCE,ART(SING){an}
:      :NPHD,CN(SING){allusion}
:      :NPPO,PP()
:      :  P,PREP(){to}
:      :  PC,NP()
```

```
:       :     DT,DTP(SING)
:       :       :DTCE,ART(SING){the}
:       :     NPPR,AJP()
:       :       :AJHD,ADJ(ABS){final}
:       :     NPHD,CN(SING){phase}
:       :     NPPO,PP()
:       :       P,PREP(){in}
:       :       PC,NP()
:       :       : DT,DTP(SING)
:       :       :   DTCE,ART(SING){the}
:       :       : NPHD,CN(SING){career}
:       :       : NPPO,PP()
:       :       :   P,PREP(){of}
:       :       :   PC,NP()
:       :       :     NPHD,PRN(SING)
:       :       :       :NOFU,NOCA(WPART){Sherlock}
:       :       :       :NOFU,NOCA(WPART){Holmes}
: PUNC,PUNCM(PER){.}
```

The analysis of passive constructions is exemplified in Figure 9. The sentence pattern found in this case is SU-VB-CS. It is looked upon as the passive counterpart of the complex transitive pattern found in active sentences, i.e. SU-VB-OD-CO. This explains the fact that the complementation type associated with the main lexical verb is given as complex transitive (CXTR), while the complementation type for the verb phrase, including the passive auxiliary, is given as intensive (INTENS).

The analyses represented in Figures 5-9 illustrate some aspects of the description of narrative style and indirect speech. In the grammar a number of rules were incorporated that were devised to also handle direct speech. These are discussed below, together with some analyses.

---

[11] Apart from the patterns presented here (which were included in the grammar and which appear to account for most direct speech, we occasionally come across some minor types. These include

- RPGU-RPDT-RPGT

For the description of direct speech the following basic patterns[11] were distinguished

- RPGU-RPDT

  the textual unit consists of a reporting utterance (RPGU), followed by a reported tail (RPDT); for example,

  (6) And on a lower note she repeated, "Robin, Robin!"
  (7) Harry said, 'I wonder why the devil did you get me into this?'

- RPDU(-RPGT)

  the textual unit consists of a reported utterance (RPDU), possibly followed by a reporting tail (RPGT); for example,

  (8) "I'm afraid not," Appleby said at once.
  (9) "What's that?"

- DSRP-RPGI-DSRP

  the textual unit consists of a discontinuous report (DSRP), followed by a reporting insert (RPGI), followed by a discontinuous report (DSRP); for instance

  (10) "Solo," he said gently, "wake up."
  (11) "You can have a word with this William yourself, Hoobin," he said, "and judge whether he seems sober and reliable.

Two remarks are in order here. The first concerns the analysis of discontinuous direct speech. No attempt is made to analyze the direct speech as one (albeit discontinuous) constituent. In other words, the analysis of each of the two parts runs autonomously. This is done for the following reasons: (1) it is in line with the way other discontinuous structures are accounted for (see below); and (2) the coherence between the discontinuous parts does not always exist, nor is it always possible to analyze them as one constituent; for example,

---

- DSRP-RPGI-DSRP-RPGI-DSRP

So far these have not been incorporated in the grammar.

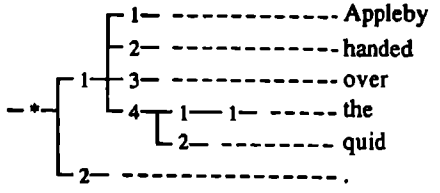
(12) 'Do', he hesitated, 'do you remember anything?'
(13) "A double funeral," Brennan said happily, "and the fucking roses wilting all over the place."

The analysis of the text found in excerpt 2 presents no problems to the grammar. The analyses obtained are represented in Figures (10)-(16). At this point we refrain from including the rather trivial analyses of the paragraph marker and the dialogue indentation markers.

excerpt 2

|Appleby handed over the quid. To add a second would, he decided, be an act of possibly offensive benevolence. He then realized that his zeal in thinking up a career for William Lockett had caused him to pass over a point of some interest in their discussion.
&"By the way,**" he said, *"why should my suggesting that Mr. Carson might take you on full-time strike you as funny? It seems a perfectly reasonable idea to me *- particularly as you and your father are used to working together.**"
&"What do you know? The man's clean busted *- just as much as this bloody service station.**"

Figure 10: Appleby handed over the quid.

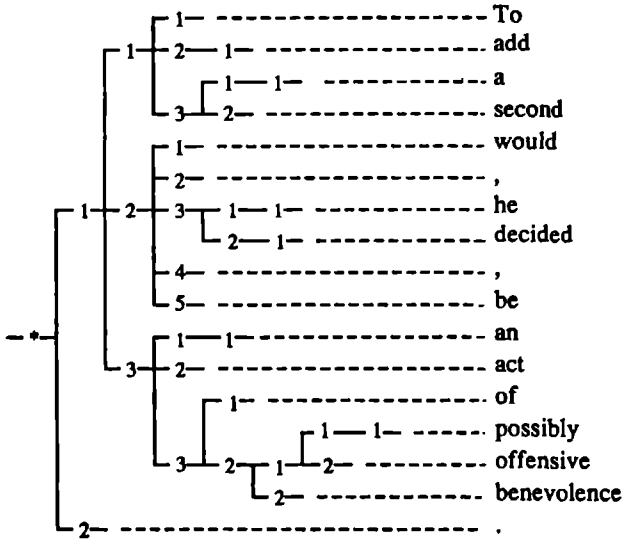



```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   SU,NP()
:     NPHD,PRN(SING){Appleby}
:   VB,VP(MOTR)
:     MVB,MLV(MOTR,PAST){handed}
:   A,AVP(PHRAS)
:     AVHD,ADV(PHRAS){over}
:   OD,NP()
:     DT,DTP(SING)
```

```
:       :DTCE,ART(SING){the}
:     NPHD,CN(SING){quid}
: PUNC,PUNCM(PER){.}
```

The analysis represented in Figure 10 is straightforward. The sentence pattern here is basically SU-VB-OD. The analysis exemplifies the analysis of phrasal verbs.

Figure 11: To add a second would, he decided, etc.



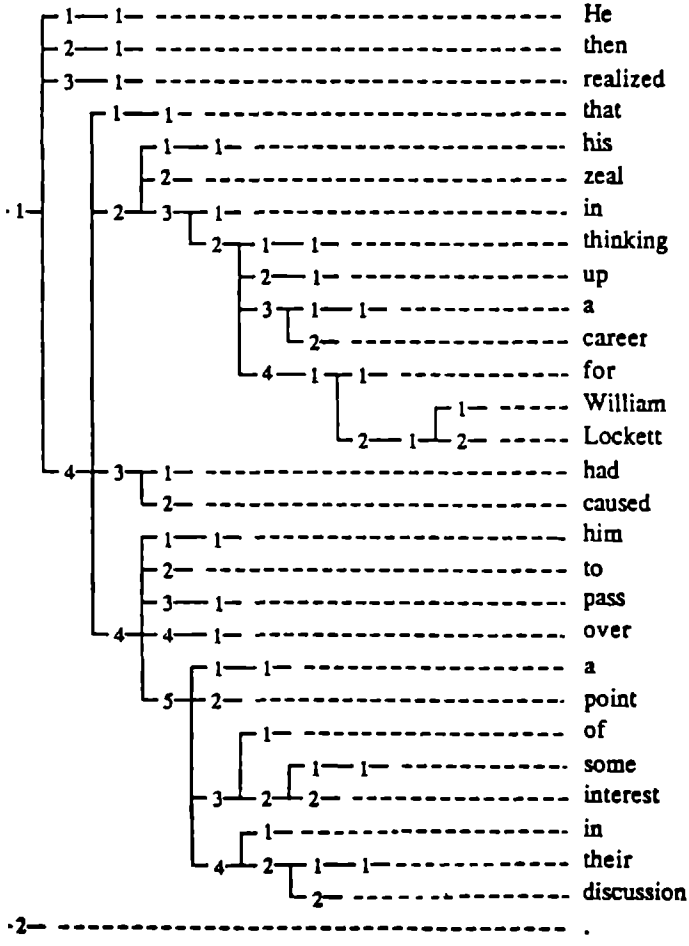

```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   SU,NFC()
:     PART,PRTCL(TO){To}
:     VB,VP(MOTR)
:       :MVB,MLV(MOTR,INFIN){add}
:     OD,NP()
:       :DT,DTP(SING)
:       :   DTCE,ART(SING){a}
:       :NPHD,ORD(SING){second}
:   VB,VP(INTENS)
```

```
:      AVB,AUX(MODAL,PAST){would}
:      NOFU,NOCA(IGN){,}
:      A,PCL()
:        :SU,NP()
:        :   NPHD,PN(PERS,SING){he}
:        :VB,VP(INTR)
:        :   MVB,MLV(INTR,PAST){decided}
:      NOFU,NOCA(IGN){,}
:      MVB,MLV(INTENS,INFIN){be}
:    CS,NP()
:      DT,DTP(SING)
:        :DTCE,ART(SING){an}
:      NPHD,CN(SING){act}
:      NPPO,PP()
:        :P,PREP(){of}
:        :PC,NP()
:        :   NPPR,AJP()
:        :      AJPR,AVP(GE)
:        :         AVHD,ADV(GE,ABS){possibly}
:        :      AJHD,ADJ(ABS){offensive}
:        :   NPHD,CN(MASS){benevolence}
: PUNC,PUNCM(PER){.}
```

There are two elements in the analysis represented in Figure 11 that deserve our attention. First, the subject is realized by a non-finite clause. While in the majority of sentences the subject is realized by a noun phrase, we occasionally come across subjects realized by finite or non-finite clauses. The description of such instances poses a problem since any such description is likely to be left-recursive, in which case it cannot be handled by the parser. Moreover, our experiences with clauses in other functions have demonstrated that embedded clauses tend to increase the complexity of the parser considerably, so that parsing times will increase accordingly. In the light of the observed (low) frequency of subjects that are realized by a clause, we decided to restrict the application of the rules describing such instances, by making the provision that the rules can only be triggered by the linguist through an intervention. The second point of interest in the analysis above is the parenthetic clause (PCL) which occurs as an adverbial in the verb phrase. The clause is looked upon as a parenthetic clause since it interrupts the main clause. Unlike a finite clause in the function of adverbial, a parenthetic clause need not be introduced by an overt subordinator.

analysis in Figure 12 is particularly interesting because of its dded clauses. The sentence pattern here is SU-A-VB-OD. The t object is realized by a finite clause. Within this clause again a transitive pattern is found. Embedded we find a non-finite (*to*-itive) clause as direct object, while also the prepositional comple- in the PP functioning as postmodifier in the subject NP is a non- clause (*-ing* participle clause).

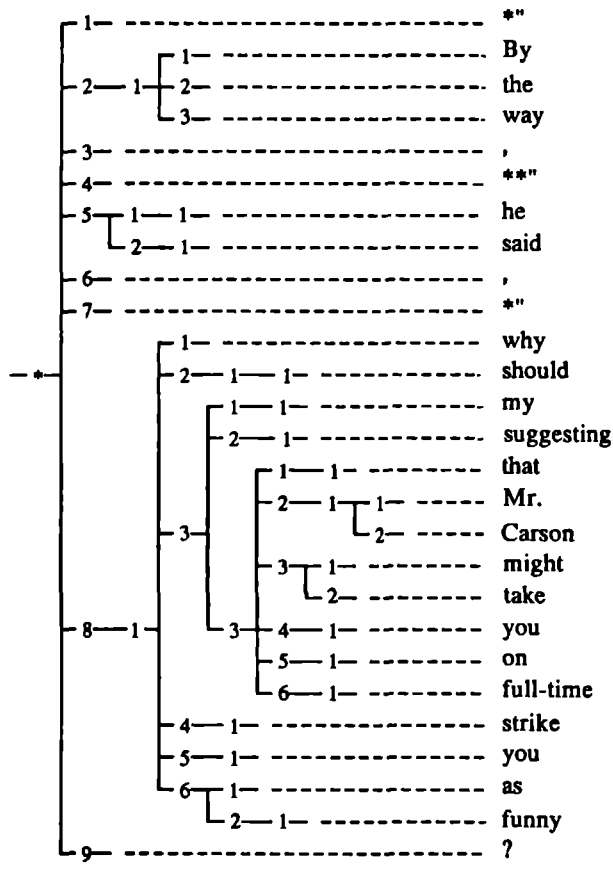
e 12: He then realized that his zeal etc.



```
NOFU,TXTU()
: UTT,S(REG,DECL)
:   SU,NP()
:     NPHD,PN(PERS,SING){He}
:   A,AVP(GE)
:     AVHD,ADV(GE,ABS){then}
:   VB,VP(MOTR)
:     MVB,MLV(MOTR,PAST){realized}
:   OD,FC(SUBORD)
:     SUB,SUBP()
:       :SUBHD,COSU(){that}
:     SU,NP()
:       :DT,DTP(MASS)
:       :   DTCE,DET(POSS,MASS){his}
:       :NPHD,CN(MASS){zeal}
:       :NPPO,PP()
:       :   P,PREP(){in}
:       :   PC,NFC()
:       :     VB,VP(MOTR)
:       :       MVB,MLV(MOTR,PRESP){thinking}
:       :     A,AVP(PREP)
:       :       AVHD,ADV(PREP){up}
:       :     OD,NP()
:       :       DT,DTP(SING)
:       :       : DTCE,ART(SING){a}
:       :       NPHD,CN(SING){career}
:       :     A,PP()
:       :       P,PREP(){for}
:       :       PC,NP()
:       :       : NPHD,PRN(SING)
:       :       :   NOFU,NOCA(WPART){William}
:       :       :   NOFU,NOCA(WPART){Lockett}
:     VB,VP(MOTR)
:       :AVB,AUX(PERF,PAST){had}
:       :MVB,MLV(MOTR,PASTP){caused}
:     OD,NFC()
:       :SU,NP()
:       :   NPHD,PN(PERS,SING){him}
:       :PART,PRTCL(TO){to}
:       :VB,VP(MOTR)
:       :   MVB,MLV(MOTR,INFIN){pass}
```

```
:       :A,AVP(PREP)
:       :  AVHD,ADV(PREP){over}
:       :OD,NP()
:       :  DT,DTP(SING)
:       :    DTCE,ART(SING){a}
:       :  NPHD,CN(SING){point}
:       :  NPPO,PP()
:       :    P,PREP(){of}
:       :    PC,NP()
:       :      DT,DTP(MASS)
:       :      : DTCE,DET(ASS,MASS){some}
:       :      NPHD,CN(MASS){interest}
:       :  NPPO,PP()
:       :    P,PREP(){in}
:       :    PC,NP()
:       :      DT,DTP(SING)
:       :      : DTCE,DET(POSS,SING){their}
:       :      NPHD,CN(SING){discussion}
: PUNC,PUNCM(PER){.}
```

Figure 13: *"By the way,**" he said, *"why should etc.




```
NOFU,TXTU()
: PUNC,PUNCM(OQUOD){*"}
: DSRP,RPDS()
:   ADCO,CON()
:     NOFU,NOCA(WPART){By}
:     NOFU,NOCA(WPART){the}
:     NOFU,NOCA(WPART){way}
: PUNC,PUNCM(COM){,}
: PUNC,PUNCM(CQUOD){**"}
```

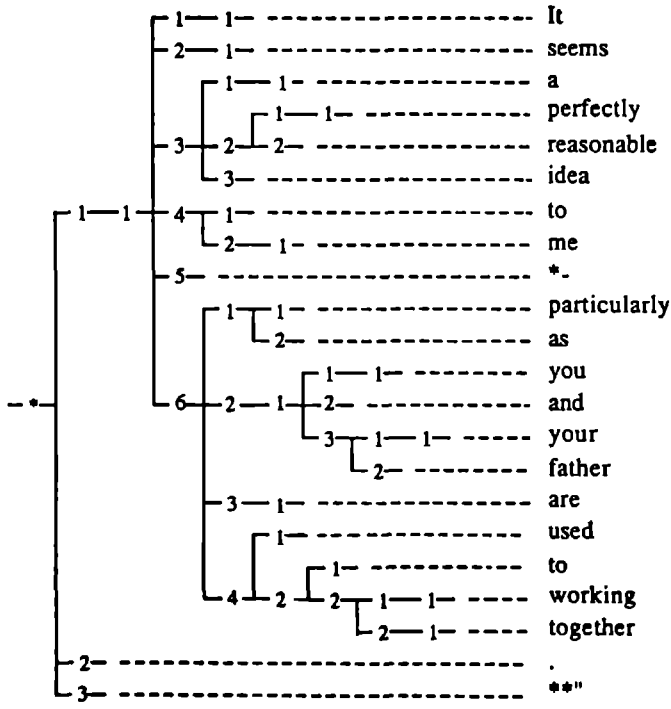
```
: RPGI,S(REG,DECL)
:   SU,NP()
:     NPHD,PN(PERS,SING){he}
:   VB,VP(INTR)
:     MVB,MLV(INTR,PAST){said}
: NOFU,NOCA(IGN){,}
: PUNC,PUNCM(OQUOD){*"}
: DSRP,RPDS()
:   CM,S(REG,INTER)
:     A,AVP(INTER)
:       :AVHD,ADV(INTER){why}
:     PROP,PROPP()
:       :OP,OPP()
:       :   AVB,AUX(MODAL,PAST){should}
:     SU,NFC()
:       :SU,NP()
:       :   NPHD,PN(POSS,SING){my}
:       :VB,VP(MOTR)
:       :   MVB,MLV(MOTR,PRESP){suggesting}
:       :OD,FC(SUBORD)
:       :   SUB,SUBP()
:       :     SUBHD,COSU(){that}
:       :   SU,NP()
:       :     NPHD,PRN(SING)
:       :       NOFU,NOCA(WPART){Mr.}
:       :       NOFU,NOCA(WPART){Carson}
:       :   VB,VP(CXTR)
:       :     AVB,AUX(MODAL,PAST){might}
:       :     MVB,MLV(CXTR,INFIN){take}
:       :   OD,NP()
:       :     NPHD,PN(PERS,SING){you}
:       :   A,AVP(PREP)
:       :     AVHD,ADV(PREP){on}
:       :   CO,AJP()
:       :     AJHD,ADJ(ABS){full-time}
:     VB,VP(MOTR)
:       :MVB,MLV(MOTR,INFIN){strike}
:     OD,NP()
:       :NPHD,PN(PERS,SING){you}
:     A,PP()
:       :P,PREP(){as}
```

```
:        :PC,AJP()
:        :  AJHD,ADJ(ABS){funny}
: PUNC,PUNCM(QM){?}
```

The analysis of discontinuous direct speech is exemplified in Figure 13. Here we find a textual unit realized by a discontinuous report (DSRP), followed by a reporting insert (RPGI) and a discontinuous report. The first discontinuous report is realized by a reported string (RPDS) which consists of a single element, viz. a connective adjunct (ADCO), which is realized by the compound connective (CON) *by the way*. The reporting insert is a regular declarative sentence. The second discontinuous report, like the first, is realized by a reported string. The sole constituent is the communicated message (CM), which is realized by a regular interrogative sentence. Note that the auxiliary verb here acts as operator (OP). In the analysis represented here the subject of the interrogative sentence is taken to be realized by a non-finite clause. It must be observed that this is only one of two possible analyses. The other analysis consists in assigning *my* the word class tag "DET" and *suggesting* the word class tag "CN". Together with the finite clause (*that Mr. Carson might take you on full-time*), the determiner and the common noun would yield an analysis as NP.

Another example of the analysis of direct speech can be found in Figure 14. Unlike the direct speech in Figure 13, there is no discontinuity. The analysis is straightforward. Note the analysis of the coordinated NPs as subject in the adverbial clause (*you and your father*).

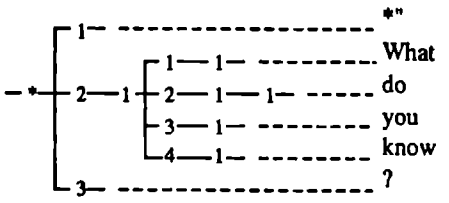
Figure 14: It seems a perfectly reasonable idea etc.




```
NOFU,TXTU()
: RPDU,RPDS()
:   CM,S(REG,DECL)
:     SU,NP()
:       :NPHD,PN(PERS,SING){It}
:     VB,VP(INTENS)
:       :MVB,MLV(INTENS,PRES){seems}
:     CS,NP()
:       :DT,DTP(SING)
:       :   DTCE,ART(SING){a}
:       :NPPR,AJP()
:       :   AJPR,AVP(GE)
:       :     AVHD,ADV(GE,ABS){perfectly}
:       :   AJHD,ADJ(ABS){reasonable}
```

```
:       :NPHD,CN(SING){idea}
:      A,PP()
:       :P,PREP(){to}
:       :PC,NP()
:       :  NPHD,PN(PERS,SING){me}
:      NOFU,NOCA(IGN){*-}
:      A,FC(SUBORD)
:       :SUB,SUBP()
:       :  SUBMO,AVP(GE)
:       :    AVHD,ADV(GE,ABS){particularly}
:       :  SUBHD,COSU(){as}
:       :SU,NOCA()
:       :  NOFU,NOCA(COORD)
:       :    CJ,NP()
:       :      NPHD,PN(PERS,SING){you}
:       :    COOR,COCO(){and}
:       :    CJ,NP()
:       :      DT,DTP(SING)
:       :       : DTCE,DET(POSS,SING){your}
:       :      NPHD,CN(SING){father}
:       :VB,VP(INTENS)
:       :  MVB,MLV(INTENS,PRES){are}
:       :CS,AJP()
:       :  AJHD,ADJ(ABS){used}
:       :  AJPO,PP()
:       :    P,PREP(){to}
:       :    PC,NFC()
:       :      VB,VP(INTR)
:       :       : MVB,MLV(INTR,PRESP){working}
:       :      A,AVP(GE)
:       :       : AVHD,ADV(GE,ABS){together}
: PUNC,PUNCM(PER){.}
: PUNC,PUNCM(CQUOD){**"}
```

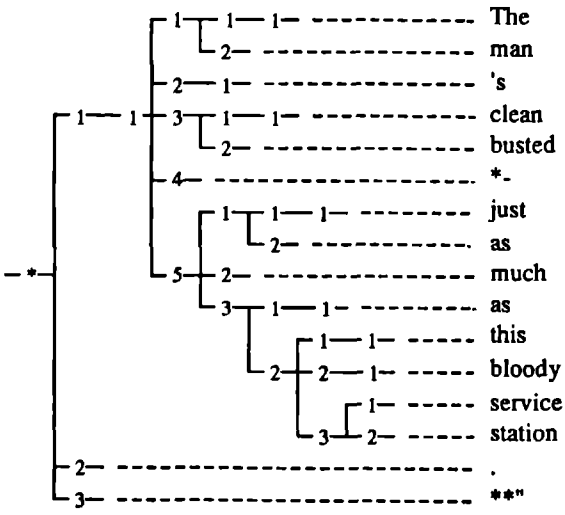
Figure 15: *"What do you know?




```
NOFU,TXTU()
: PUNC,PUNCM(OQUOD){*"}
: RPDU,RPDS()
:   CM,S(REG,INTER)
:     OD,NP()
:       :NPHD,PN(INTER,SING){What}
:     PROP,PROPP()
:       :OP,OPP()
:       :   AVB,AUX(DO,PRES){do}
:     SU,NP()
:       :NPHD,PN(PERS,SING){you}
:     VB,VP(MOTR)
:       :MVB,MLV(MOTR,INFIN){know}
: PUNC,PUNCM(QM){?}
```

In Figure 15 we find an example of a regular interrogative sentence in which the direct object occurs sentence initially. Finally, in Figure 16 we find once more a reduced finite clause as adverb phrase postmodifier (AVPO).

Figure 16: The man's clean busted *- just as etc.




```
NOFU,TXTU()
: RPDU,RPDS()
:   CM,S(REG,DECL)
:     SU,NP()
:       :DT,DTP(SING)
:       :   DTCE,ART(SING){The}
:       :NPHD,CN(SING){man}
:     VB,VP(INTENS)
:       :MVB,MLV(INTENS,ENCL,PRES){'s}
:     CS,AJP()
:       :AJPR,AVP(GE)
:       :   AVHD,AV(GE,ABS){clean}
:       :AJHD,ADJ(ABS){busted}
:     NOFU,NOCA(IGN){*-}
:     A,AVP(GE)
:       :AVPR,AVP(IN)
:       :   AVPR,AVP(GE)
:       :     AVHD,ADV(GE,ABS){just}
:       :   AVHD,ADV(IN,ABS){as}
:       :AVHD,ADV(GE,ABS){much}
```

```
:       :AVPO,FC(RED)
:       :  SUB,SUBP()
:       :    SUBHD,COSU()[as]
:       :  SU,NP()
:       :    DT,DTP(SING)
:       :      DTCE,DET(DEM,SING)[this]
:       :    NPPR,AJP()
:       :      AJHD,ADJ(ABS)[bloody]
:       :    NPHD,CN(SING)
:       :      NOFU,NOCA(WPART)[service]
:       :      NOFU,NOCA(WPART)[station]
: PUNC,PUNCM(PER)[.]
: PUNC,PUNCM(CQUOD)["]
```

The grammar describes most syntactic structures, unmarked as well as marked ones. Marked structures may be of two kinds: (1) they are either straightforward permutations of the basic sentence or clause patterns, i.e. they differ from the structures described by the basic patterns (see section 4.2) in that they show a deviant word order, or (2) they contain some provisional element, such as provisional *it* or existential *there*. While for the latter group of structures the surface syntactic description, which in the handbooks on English grammar remains implicit, had to be made explicit, the analysis of these structures is not all that problematic. The former, on the other hand, constitute a problem since the description of all sorts of deviant word orders brings about a certain amount of indeterminacy with respect to the distinction of particular constituents and the functional relations that hold between these.

The grammar can be said to be fairly complete. It not only describes structures in which unmarked word order is found, but also such structures as cleft, existential and extraposed sentences, verbless clauses, interrogatives, imperatives, clauses or sentences in which subject-verb inversion occurs, etc. Furthermore, a description is provided for instances of direct speech, which includes certain marked and highly elliptic clause structures, as well as some typical discourse elements (e.g. formulaic expressions, forms of address, connectives, interjections), enclitic forms, etc. Note that when the grammar is said to be fairly complete, 'complete' has a relative meaning: the grammar can only be considered to be complete with respect to the description of the standard variety of the language. As a consequence, when in the analy-

sis process we come across text which originates from another variety, the grammar falls short and analysis fails. This is for instance the case with some of the text encountered in excerpt 3, i.e. a passage in which Hoobin, gardener on the Appleby estate, speaks his mind.

excerpt 3

> *&*"And fair's fair,**" Hoobin went on, unheeding. *"It mayn't be the Carson body's fault, for all I know. Not with all them crooks and cheats in stock exchanges and such like places. What with it coming all that bad, the lad William deserves his chance. I'll try him, that I will. For I'm an open-minded man, I am. And it's the perusing does it, Sir John.**" Upon this elevating thought, Hoobin picked up his *2{Daily Mirror} *0again. *"There do seem to have been a terrible great child murder in Houndsditch,**" he said. *"There be a whole column on it. I've been reading about it, I have, this half hour and more.**"

5.3 An (informal) assessment of the grammar and its performance

In the preceding section an outline of the analysis process was presented, together with some analysis results that were obtained. An evaluation of the analysis results shows us that the overall performance of the grammar in terms of coverage is quite satisfactory. Only a small number of constructions have not yet been accounted for in the grammar. As a consequence, the analysis of an utterance fails occasionally. On the whole, this happens in the case of constructions (often minor variants of common constructions) which occur relatively low-frequently. Their description is likely to be incorporated in the grammar in the near future, when we have gained sufficient insights as to their nature, their form and distribution of occurrence. In the handbooks on English grammar the description of these constructions often remains implicit or has been omitted altogether.

Below a number of structures and phenomena are discussed in the light of the grammar that was developed and implemented. Special attention is given to the problems that were encountered in the description of some of these structures, and the solutions that were arrived at.

Transposition

The term *transposition* was introduced in section 3.5 as an umbrella term for all phenomena that involve fronting, inversion or postponement of (functional) constituents. Using the umbrella term, various forms of transposition that in the literature are described under a variety of names can be classified employing the following three factors:

1. the direction of the transposition (forward or backward);

2. the distance over which the transposition takes place (contiguous, i.e. to a position immediately preceding or following the constituent that is adjacent to the current constituent, or non-contiguous, i.e. to a position more than one constituent removed from the current position);

3. the constituents involved (ICs of sentences or clauses, ICs of phrases, or words).

On sentence or clause level we find instances of forward and instances of backward transposition. Each of these is exemplified below. Note that with forward transposition we only find non-contiguous transposition. The use of an explicit provisional element or operator here prohibits contiguous transposition. With backward transposition both contiguous and non-contiguous transposition occurs.

- non-contiguous forward transposition; this kind of transposition requires the presence of an explicit operator, e.g. *it* in cleft-sentences, and *there* in existential sentences. For example,

  (14) It is presumably the former which suggests the possibility of some changes in market power.
  (15) There is so much to look forward to.

- contiguous backward transposition; traditionally, the literature refers to this phenomenon as inversion. The most common types of inversion are those of subject-verb inversion (16)-(17) and subject-operator inversion (18)-(19), for example,

  (16) To each ex-works price was added a transport charge for each five miles, the charge being increased less than in proportion to distance.
  (17) "And no more have I any business to be meddling," he said.

(18) Thus did Sir John Appleby, a citizen tolerably well seen in human nature, meditate dispassionately on the Carsons of Garford House.
(19) Not only did it not contain Carys, it contained no trace of her existence there.

A less common type occurs in the complementation structure, where the object complement may precede the direct object:

(20) The notion of the Carsons, father and son, as being jointly engaged in financial and other dispositions, making possible a simultaneous flight from the tiresomely dotty Cynthia Carson, mightn't stand up to scrutiny, yet it held a certain attractiveness for the speculative mind.
(21) Early on, Hackett had discarded as purposeful fictions most of the titbits James had let drop about himself.

- non-contiguous backward transposition; backward, non-contiguous transposition is the fronting or preposing of objects or complements. Such transposition occurs with active as well as with passive sentences. As a result, apart from having to describe the basic sentence patterns (cf. section 4.2) and their passive equivalents, we also have to include in our grammar their permutations, i.e. we have to provide the description of such patterns as listed below (n.b. the related basic clause patterns are given between brackets).

CS - SU - VB	(SU - VB - CS)
OD - SU - VB	(SU - VB - OD)
OI - SU - VB - OD OD - SU - VB - OI	(SU - VB - OI - OD)
OD - SU - VB - CO CO - SU - VB - OD	(SU - VB - OD - CO)
OI - SU - VB - OD - CO OD - SU - VB - OI - CO CO - SU - VB - OI - OD	(SU - VB - OI - OD - CO)

Also on phrase level we find instances of forward and backward transposition. As with immediate constituents of sentences or clauses, forward transposition of ICs of phrases is always non-contiguous, while backward transposition may either be contiguous or non-contiguous. Each of these is exemplified below.

- non-contiguous forward transposition; the most common type of forward transposition in the case of phrases is usually referred to as

postponement (cf. Quirk et al., 1985). For instance, with NPs (22), AJPs (23), AVPs (24) and DTPs (determiner phrases, 25) the postmodifier need not immediately follow the head it modifies but may be postponed. For example,

(22) The time had come *to lay his cards on the table.*
(23) Mentalistic hypothetical constructs have a different kind of ontological status *from that of physiological hypothetical constructs.*
(24) In psychology, however, the concept of cause is far more complex *than in other disciplines.*
(25) Not since the halcyon days of his big wins at the casinos had Marty possessed so much money *as he did now.*

With prepositional phrases the preposition may be postposed.[12] Consider the following examples:

(26) The book that the article was published *in* was no longer available.
(27) What performance did he go *to*?

Other instances of non-contiguous forward transposition are found with appositives (floating apposition) and determiners (deferred determiners).

- contiguous backward transposition; this type of transposition is found in NPs where the premodifier may precede the determiner. For example,

(28) quite as dismissive a word
(29) too urgent a matter

- non-contiguous backward transposition; this type of transposition appears to be extremely rare in the case of phrases. Occasionally, however, we come across postmodifiers that have been fronted and occur (probably exclusively) in sentence or clause initial position. For example,

(30) *Than these last five words* he positively felt that he had never heard anything odder in life.

[12] Note that, as Quirk (1972: 396, note a) points out, in some instances the postposed preposition has no preposed alternative. This holds true for a sentence like *What did you do that for?*

The description of instances of transposition as exemplified above is, from a descriptive point of view, straightforward and unproblematic. The grammar can simply be extended with alternative rules so as to account for various deviant word orders. However, as observed earlier (section 5.2), with the inclusion of rules describing other, deviant word orders, the distinction of constituents and their functional roles becomes subject to an increasing amount of indeterminacy, causing the analysis to be more ambiguous. This can easily be illustrated by looking at the effect that the inclusion of a single rule may have. Consider the case of the contiguous backward transposition found with the complementation structure (cf. examples 20-21). At first our grammar described merely the unmarked word order found in complex transitive sentences, that is, the order in which the direct object and the object complement occur is such that the direct object precedes the object complement. The analysis of a sentence like (31) yields a non-ambiguous result: the structure assigned to this sentence is SU-VB-OD-CO.

(31) They recommended him as editor-in-chief.

The inclusion in the grammar of a rule describing the transposition found in (20)-(21), thus allowing for the object complement to precede the direct object, causes the analysis of (31) to be ambiguous. Apart from the analysis we had before, a second analysis is obtained in which the structure that is assigned is SU-VB-CO-OD.

The above shows the effect that a simple extension of the grammar may have. Although the scope of the current grammar is restricted to the syntax of the language, so that we cannot depend on semantic information to come to grips with any ambiguity that might thus arise, it must be held feasible to formulate sets of restrictions that make it possible to control this effect to some extent. Handbooks on English grammar such as Quirk et al. offer little help in this respect. For example, Quirk et al.'s description of subordinate clauses constitutes a relatively clear case of how a rather extensive (unformalized) description fails to make explicit the criteria that can be used to identify a particular structure. Looking for criteria that will help us to identify such zero subordinate clauses as exemplified in (32)-(34), we cannot but conclude that the information Quirk et al. provide cannot possibly be operationalized in a formal grammar.

(32) '*Had I been allowed*, I would have nurtured you from childhood.'

(33) *Had it, however, acquired its original dominating position by an amalgamation of smelters*, then, in Judge Hand's view, that would have provided clear grounds for a charge.
(34) Judith would have to be a person of quite morbid sensibility *were she to be thrown into a state of distress by the nonarrival at his parents' dwelling of a totally strange young man.*

For instance, discussing the theme in subordinate clauses, Quirk et al. (1972: 950) observe that

> "In subordinate clauses, the usual thematic elements are subordinators, *wh*-elements, and the relative pronoun *that*. Special frontings of other elements as theme occur only in idiomatic or literary constructions of minor importance ..."

In their examples they present a wide variety of what they call "unusual syntactic orderings" (Quirk et al., 1972: 749). Their characterization of these structures is as follows:

> "A device which may replace the subordinator *if* in signalling a conditional clause is the inversion of subject and operator, particularly with the operator *had* in hypothetical clauses ..."
>
> Quirk et al. (1972: 748)

> "Subjunctive *were* and hypothetical or putative *should* can also undergo inversion in somewhat literary style ..."
>
> Quirk et al. (1972: 748)

> "Like conditional clauses, concessive clauses sometimes have unusual syntactic orderings. The subordinators *as*, *though*, and *that* occur in non-initial position after the subject complement ... *That* and *as*, in this position, can also have the non-concessive meaning of cause or circumstance ... The rule which permits this construction applies more generally to *as* and *though*, such that a whole predication (consisting *eg*, of lexical verb, or lexical verb plus object) may be placed in front of the conjunction: *object as you may; fail though I did; change your mind as you will.* In *much as you like to help*, on the other hand, it is an adverb alone that is fronted. Such clauses, rather formal in style, may be compared with conditional-concessive clauses such as *come what may* ..."
>
> Quirk et al. (1972: 749-750)

It will be clear that such descriptions as "in idiomatic or literary constructions of minor importance", "in hypothetical clauses", "in somewhat literary style", etc. cannot possibly be formalized. Therefore, we cannot but conclude that we have to look elsewhere for restrictive conditions, which may take the form of restrictions on the realization of particular functions in certain contexts. A study of the material that has already been analyzed might help to uncover (part of) the information we are looking for. In the case of OD-CO inversion, for instance, the analyzed material suggests that in the case of inversion the realization of the CO is restricted to AJP and PP.

Apart from the forms of transposition discussed above, there are two others that we come across. These types of transposition, exemplified in (35)-(45), add yet other dimensions to the problems encountered in the formal description of transposition. They are briefly discussed below.

Consider the following examples:

(35) John is said to be a professional.
(36) Was the man in a state of anxiety which for some reason -- perhaps a notion of proper manly behaviour -- he felt obliged to dissimulate?
(37) There was a small puzzle here -- but Appleby told himself it was a puzzle he felt no particular impulse to resolve.
(38) And the man himself was more worried than he confessed to being.
(39) "And that something you'd rather like to do yourself?"

These structures exemplify the phenomenon of what in the literature (cf. Quirk, 1985: 1202) has been termed *raising* and what in terms of our description may be referred to as *cross-transposition*, i.e. the transposition of a constituent across the boundaries of clause constituency established in the formal grammar (cf. cross-reduction, p. 196f). Our description as contained in the formal grammar fails in two respects. First, for these structures it appears impossible to maintain the constituent structure that was postulated, without having to postulate ellipsis (cf. chapter 4). Second, the elements that are raised and as a result function in a superordinate clause still constitute a functional constituent in the clause they were raised from. Strictly speaking then, raised constituents should be labelled for both their function in the superordinate clause and the clause they were raised from. Within the current descriptive framework this is impossible, for neither do we 'reconstruct' sentences into some sort of underlying structure, nor is it possible to

assign more than one function label to a constituent. For the moment a solution is found in labelling a constituent for its function in the superordinate clause, while information about its function in the subordinate clause is contained in an affix.

The second form of transposition that deviates from what we have seen so far is exemplified in examples (40)-(47). These are instances of what is commonly referred to as *topicalization.*

(40) *It* wasn't hers, *this tree.*
(41) 'And *you, John,* I tried to kill *you.*'
(42) And *that* was another thing: *the way his belly revolted if he put food into it.*
(43) *It* rattled on their plastic shrouds, *a dry rain.*
(44) *She* had been beautiful, *this one.*
(45) *A peddler, a pushcart,* he could have got *it* anywhere.

The transposition exemplified in (40)-(45) resembles the transposition found with cleft and existential sentences since here too, an operator is involved. Unlike the operator in cleft and existential sentences, however, the operator in topicalized structures of the kind discussed here is not restricted to a typical or unique item (such as *it* or *there*). Rather, the operator is one of a class of proforms or pronouns which appears to be co-referential with the NP which constitutes the topic of the sentence or clause. When we restrict ourselves to syntactic information only, however, the operator is not uniquely identifiable. So far, topicalized structures have therefore not been included in the grammar as such. In the analysis they were dealt with as instances of floating apposition.

Insertion

A phenomenon that is more or less related to transposition is that of insertion. In so far as insertion occurs at the boundaries of phrases that realize syntactic functions on the level of the sentence or clause, its description is unproblematic. For example, the parenthetic clause found in (46) presents no problems whatsoever:

(46) This, *it was argued,* led to economies in production.

Commonly, however, inserted elements in the form of parenthetic clauses, interjections and suchlike items are found *within* the bound-

aries of phrases, where they form a problem both for the description and the subsequent analysis. Consider the following examples:

(47) 'A few of us in this room now know the cause of these outbreaks; it is my intention that you all know, so that we can combine our various skills to combat this growing -- *and I mean this literally* -- threat.'
(48) 'And is this, *er*, disease infectuous?' Sir Trevor Chambers asked, carefully avoiding Holman's eyes.
(49) 'We have a meeting in,' *he looked at his watch*, 'ten minutes.'

From a descriptive point of view the main problem with insertion lies in the fact that in violating the self-contained units of description that phrases normally are, the descriptive capacity of the constituency model is seriously affected. If insertion must be assumed to occur freely, linguistic description in terms of constituency becomes impossible since it is no longer possible to identify all of the basic descriptive units.

In handbooks on English grammar insertion is described as a phenomenon that occurs, typically, with varieties that are less formal. Inserted elements often indicate hesitations, self-corrections, attempts at some clarification of what was said earlier, etc. As we have seen before (cf. transposition), it is impossible to operationalize such information in a formal description. For the time being the analysis of instances of insertion must therefore be triggered by the linguist in an intervention. Not until more systematic information becomes available about the circumstances under which insertion occurs, which makes possible a discovery of the regular patterns in its distribution, can a satisfactory description of insertion be arrived at.

Phenomena related to coordination

On the whole we have found that the multi-layered structure that was introduced to cope with phenomena that are usually solved by the postulation of ellipsis proves adequate. A number of problems remain, however. Some phenomena that occur in conjunction with coordination appear to be problematic when it comes to analyzing them. Among these are *zero coordination*, *correlative coordination*, *cross-reduction* and *neutralization*. Each of these is discussed below.

Zero coordination

Occasionally we come across instances of coordination where there is no overt coordinator, nor a punctuation mark which may function as such. Instances like these are too frequent to be looked upon as "errors", i.e. as the (repeated) erroneous omission of a punctuation mark. Moreover, they occur in different samples by different authors, and therefore cannot simply be held to be manifestations of some idiosyncratic feature. In all instances we have found so far the second conjoin of the coordination is introduced by *then*. Consider the following examples:

(50) For a moment the eyes looked at the policeman then swivelled back towards Holman.
(51) She hesitated a moment -- unsure of whether or not to go in -- then slipped down the stairs again, leaving him unwoken.
(52) The engine juddered, the train pulled at it then seemed to squeeze and strain through the cold, moist air, to concertina, to unfold and concertina again.

So far these instances have not been accounted for in the grammar. The data seem to suggest that *then* must be looked upon as a coordinating conjunction. In fact, it appears possible to replace *then* by *and*. However, whereas *then* explicitly indicates the chronological order in which events occur, this remains implicit when *and* is used.

Correlative coordination

In the process of analyzing corpus material instances of what is commonly described as correlative coordination did not always conform to the rules that we had formulated to describe this type of coordination. In our description of correlative coordination we had assumed that the correlative coordinators would introduce similar constituents, i.e. constituents realizing one and the same function. This implied that the first part of a correlative pair was expected to precede the first conjoin in the coordination immediately. However, apart from instances that did fit this description, we also came across instances such as those found in examples (53)-(57), where the first part of the correlative pair occurred closer towards the beginning of the sentence or clause. These seemed to suggest that the constituent structure we had postulated in our grammar was inadequate. On the other hand, there was much evidence in support

of this structure. A solution was found in no longer describing correlative coordinators as such; rather, the two parts of a correlative pair are analyzed independently. The first part is analyzed as an adverb realizing the function of an adverbial serving as a coordination signal, the second part is analyzed as a coordinating conjunction acting as coordinator.

(53) They entail predictability which can *either* be deterministic or statistical.
(54) Whenever one observes or infers a sequence, whether the sequence is observable or theoretical, then that sequence may *either* be a causal process or a pseudoprocess.
(55) That is, the events could (in principle) be described in terms of *either* a physiological, mentalistic or mechanistic process.
(56) The concept of cause is thus commonly used *both* for sequential and simultaneous relations.
(57) 'And remember, nothing can ever be done about the healthy cells that have been damaged *either* by the parasites or X-ray.

Note that apart from items like *either, neither* and *both* (parts of the traditional correlative coordinator pairs *either ... or, neither ... nor*, and *both ... and*), we also come across items like *whether* which occur in a similar fashion. Consider the following examples:

(58) The idea of variation or change in the form of an operator is independent of *whether* it is simple or complex.
(59) In scientific methodology the experiment (discussed in chapter 9) provides a paradigm for deciding *whether* a relation is causal or not, and it does so in terms of manipulated or marked independent variables.

Here it may be observed that *whether* can be regarded as a subordinating conjunction which occurs in the function of subordinator.

Cross-reduction

The term *cross-reduction* that is introduced here refers to instances of coordination where constituent structure appears to be violated in that coordination occurs across the levels of constituency established in the formal grammar. This phenomenon occurs most frequently at sentence or clause level, although occasionally it is found with phrases as well. Consider the following examples:

(60) Then we could analyse it, discover its contents, and then develop a serum.
(61) Could it have drifted into Winchester Cathedral and become trapped inside its ancient but solid stone walls?
(62) He was acknowledged and permission granted.
(63) 'But do we have time to experiment and develop ideas?'
(64) 'I'm going to take you downstairs and hand you over to Mrs Janet Halstead, Principal Medical Officer for the Research Council.
(65) Once the body's system has beaten off a disease it builds some, or often total, resistance against it, and in this case, where the mutated mycoplasma would be virtually flushed from the system in the early stages and the unwanted cells in the brain killed before they had a chance to form, as they have in Mr Holman's case, then, yes, I believe one could be made immune for further attacks.
(66) In this and later chapters entities as well as concepts in psychology will be examined.

Instances of cross-reduction are instances of coordination that cannot be accounted for in terms of the multi-layered constituent structure that has been established. The reader will recall that this structure replaces the flat structure that we find in for instance Quirk et al. (1972, 1985). This multi-layered structure makes it possible to cope with phenomena that otherwise must be accounted for by postulating ellipsis. For reasons that were set out and explained earlier (cf. chapter 4) a description which postulates ellipsis is considered to be unattractive. Although on the whole the multi-layered structure proves to be adequate, we find that with some instances the (current) multi-layered structure does not provide a solution. For example, in the case of example (60) we have an instance of coordination of predicates which -- even with the current multilayered structure -- requires the postulation of ellipsis of the auxiliary in the second and third conjoin. In (61) the coordination of sentences may be postulated with ellipsis of the auxiliaries *could* and *have* and the subject in the second conjoin. Slightly different is example (62) which shows coordination with ellipsis of the auxiliary in the second conjoin. Example (63) exemplifies an instance of cross-reduction with non-finite, infinitive clauses where the *to*-infinitive marker is ellipted in the second conjoin. In example (64) we find once more the coordination of two sentences with ellipsis of the subject and the auxiliary (cf. 61). And finally, in examples (65) and (66) instances of cross-reduction are found to occur with noun phrases. Note that the cross-reduction exemplified in (65)-(66) differs from that found in (60)-(64) not so much in that here the cross-reduction is applied to phrases instead of sentences or clauses, but rather in that the reduction is backward, not

forward.

Despite the fact that occasionally we do come across instances of cross-reduction as exemplified above, we have so far refrained from adapting or otherwise revising the constituent structure that we have postulated. This decision is motivated by the fact that cross-reduction does not occur all that frequently, while the constituent structure established in the formal grammar proves adequate in most other cases.[13] Moreover, adaptation of the constituent structure in the light of instances of cross-reduction such as the ones presented here is considered unattractive since it forces us to abandon descriptive notions like verb phrase. For reasons that were set out and explained earlier (cf. chapter 4) a description which postulates ellipsis is not considered a reasonable alternative.

Neutralization

Occasionally with instances of coordination (including cross-reductions) we may come across *neutralization*, i.e. the unification of two distinct (and mutually exclusive) characteristics, such as word class membership or feature values, within one and the same token. Neutralization is relatively frequently found with verb phrases, but there are indications that it occurs with other phrases as well. The examples given here concern both verb phrases and noun phrases. Thus in example (67) *was* carries both the interpretation of intensive verb and that of passive auxiliary; in example (68) *was* must be interpreted as intensive verb and as progressive auxiliary.

(67) The tiny oxygen tank on his back *was* uncomfortable but deemed necessary in case the mist became too choking.
(68) Somebody *was* very close, and yet not answering.

Neutralization also occurs in noun phrases, where especially determiners appear to be subject to it. Consider the following examples:

(69) The air was much harder to breathe in, the acidity burnt *his* nostrils and throat.

---

[13] Note that the descriptive problems encountered with instances of cross-reduction are similar to those experienced in the description of instances of cross-transposition.

(70) He had washed *his* hands and face of bloodstains and he smelt strongly of perfume.
(71) In sum an operator can be described at a descriptive level -- the level of *the* hypothetical *construct(s) whose* states are related -- and can also be described at an explanatory level -- some other theoretical level involving another kind of hypothetical construct.

In examples (69)-(71) the possessive determiner is neutralized as far as its value for the feature 'countability' is concerned, i.e. it combines the values "SING" and "PLU". In (71) the same goes for the definite article, the noun and the relative pronoun.

Neutralization is a phenomenon our current grammar fails to describe. Its description is problematic since it requires the association of otherwise alternative analyses -- for example DET(SING) and DET(PLU), or AUX(BE,PAST) and MLV(INTENS,PAST) to a token, something the descriptive framework is not equipped to do.

Juxtaposition

Another phenomenon we have not incorporated in the current grammar is that of *juxtaposition.* Instances of juxtaposition can be characterized as follows. Two constituents are juxtaposed when they occur as adjacent constituents, and the relation that holds between these constituents is appositive-like, in the sense that the second constituent gives a further specification of the constituent it is juxtaposed to. However, unlike occurrences that in the grammar were described as apposition, which were restricted to appositive noun phrases, instances of juxtaposition involve different categorial constituents, one of which is a noun phrase. The description of instances of juxtaposition as exemplified in (72)-(77) is problematic since it appears impossible to (automatically) identify the constituent that the noun phrase is juxtaposed to. The constituent that the noun phrase is juxtaposed to is commonly a finite or non-finite clause as in (73) and (75), although, as the other examples show, it is also possible to have a predicate or a main clause.

(72) Breer looked straight at Mamoulian, *something he very rarely mustered the courage to do.*
(73) These authors suggest that behaviour can be explained by reference to the thoughts of the person engaged in that behaviour, *a suggestion which forms one of the assumptions of modern attribution theory.*

(74) He was, of course, rejected amid much regimental laughter, *a fact he attributed to Frank Richards having made such a fool of him in print.*
(75) The method we've been using today, all day, is sprinkling calcium chloride from low flying aircraft, *a practice used in San Fransisco regularly to clear their fogs.*
(76) They kept the illegal gun hidden behind the transmit unit, *a secret agreement among themselves and many other aeroplane crews,* as a protection against the increasingly frequent hijackings.
(77) Holman felt himself almost mesmerized by her words and began to relax, *a combination of his own tiredness, the soft chair he was sitting in and the easy manner in which she was talking to him.*

As the examples show, the noun phrase is always headed by an abstract noun (such as *suggestion, fact, practice, agreement, combination* in the examples above) and postmodified by a prepositional phrase or a clause. The function of determiner is realized by the indefinite article.

5.4 Towards a standard for assessing the grammar

Up to this point our evaluation of the grammar and its performance has been rather informal. While the analyses that were included in section 5.2 served to give the reader an impression of what the analyses look like, in section 5.3 we discussed a number of structures and phenomena in the light of the problems that were encountered in the description of some of the structures, and the solutions that were arrived at. It is quite evident, however, that a more formal evaluation is required. Not only will such an evaluation give better insight as to what results were obtained and how these results were achieved, it will also present a more accurate account of the performance of the grammar and the associated parser, inform us of the reliability of the results, and -- equally important -- it will make possible a comparison with other grammars and parsers. Before an evaluation of this kind can be carried out a standard must be adopted which should provide us with objective criteria that may be used in assessing the grammar and the parser. The present section addresses the issue of what standard should be set, and gives an evaluation of the TOSCA grammar and parser in terms of the proposed criteria and measures.

5.4.1 The grammar: linguistic object and analysis tool

Before proceeding towards a discussion of what criteria and measures should be employed, it should be pointed out that an evaluation of the grammar and the parser cannot be detached from the background against which these were developed. As was observed in chapter 3, the design of the grammar is to a large extent influenced by the goals that are set. In the Nijmegen approach, for example, where the use of a grammar-based parser is preferred to that of a hard-coded one, the role attributed to the formal grammar is two-fold: it is a means for producing databases and it allows for the testing of the linguistic hypotheses incorporated in the grammar. From these views and the decision to give priority -- for the time being -- to the creation of syntactic databases certain conflicts of interest arise. These make it impossible to hold it a feasible proposition to carry out a single, overall evaluation which takes into account this double role of the grammar. Rather, separate evaluations must be made of what in the remainder of this section we shall refer to as the grammar, i.e. the grammar as a linguistic object in its own right, and the parser, i.e. the grammar as an analysis tool. While the latter is amenable to the application of formal criteria, it is our contention that the former is not.

Below we briefly review the requirements that must be made with respect to the grammar on the one hand and the parser on the other.[14] Next, a summary is given of the criteria that may be applied in evaluating the grammar and parser respectively.

I. The grammar

The grammar in its capacity of formalized description of the language, incorporating the linguistic hypotheses held by the linguist, constitutes a means for testing these hypotheses against authentic data contained in a corpus. The corpus in this role functions as a test bed, making it possible to pursue the large-scale testing of linguistic hypotheses. Where, as in the Nijmegen approach, the view is held that the formal grammar should be a comprehensive grammar aimed at the complete description

[14] A more extensive discussion on the present role of the grammar and the parser in the Nijmegen approach, and the requirements that were made, can be found in chapter 3 (more specifically in sections 3.2 and 3.3).

of the language in question (rather than at the development of yet another toy-grammar), the formalization of the hypotheses and the subsequent, extensive testing against corpus data form necessary steps in the development of robust linguistic descriptive theories. The process of formalization forces the linguist to be fully explicit about each and every aspect of his hypothesis. The testing brings to light any lacunae and/or other imperfections the hypothesis may contain. Once the grammar has been altered in view of the testing results and further testing proves it to be satisfactory, the grammar supposedly has reached the point of near exhaustiveness.

Apart from the requirements of robustness, explicitness, and exhaustiveness that were mentioned above and which belong to the domain of every empirical discipline,[15] there is another set of requirements which relate more specifically to the linguistic felicity of the grammar. For a grammar to be linguistically felicitous it must conform to the methodological standard set by the discipline. This means that the grammar must provide an empirically-based description of the structure of the language (cf. Hudson, 1981: 335) while employing the tools and methods that have acquired widespread recognition in the linguistic community. This involves the use of a (linguistic) "metalanguage containing technical terms denoting analytical categories and constructs" (Hudson, 1981: 335), and the use of a (type of) formalism that is familiar to linguists in their discipline. Moreover, the structures that are associated with the utterances of the language on the basis of the grammar must correspond with the interpretations that the language user associates with them, i.e. they must be semantically and pragmatically relevant.

While the requirements mentioned above constitute criteria for evaluating a grammar, it should be observed that they do not provide absolute measures. Rather, they must be interpreted in the light of the state of the art in the discipline at a particular time, as well as the goals that are set and the restrictions that hold. In descriptive linguistics we have not yet achieved the point where a comprehensive (formal) grammar containing a complete description of the language is a feasible proposition. The present goals are much more moderate. For example, as was out-

[15] Note that other requirements might have been listed, such as economy and elegance. Unlike the requirements mentioned earlier, however, these are considered to be of secondary importance.

lined in chapter 3, the present objective in the Nijmegen approach is to construct a formal grammar that describes the morpho-syntactic structure of individual utterances. For the time being the incorporation in the grammar of substantial semantic and pragmatic components is not at issue. In view of this the requirement of exhaustiveness of description must be translated into the exhaustive description of the syntax of the language. The explicitness of description is warranted by the use of a formal grammar.

While the criteria of explicitness and exhaustiveness can relatively easily be implemented in order to arrive at a more formal evaluation of the grammar, the criterion of linguistic felicity of description proves too elusive a concept. The main problem in implementing this criterion lies in the fact that as linguistic knowledge is extended, grammars develop and descriptions are improved upon. Standard descriptions that have emerged from the Great Tradition (see section 1.3) have yielded comprehensive grammars describing the syntax of the language. Taking these as the starting-point in writing a formal grammar we stay close to what is traditional and familiar in descriptive linguistics. However, as becomes apparent in formalizing (aspects of) these descriptions, they are incomplete and from time to time internally inconsistent or ambiguous. The descriptive framework must be extended and adapted so as to include notions that did not occur before. In so far as our grammar constitutes a formalization of established descriptive theory it can be evaluated by comparing the different versions that are or may be based on the common descriptive principles embodied in a handbook of English grammar. However, on points where the grammar provides a description where earlier descriptions failed to do so, were internally contradictory or ambiguous, no such standard for comparison is available and an evaluation can be no more than an account based on subjective judgment.

II. The parser

The parser, i.e. the grammar in its role of production tool employed as a means of creating databases, should meet the following requirements:

- it must be efficient;
- it must have full coverage of the language described;

- the analyses it produces must be consistent;
- the structures that are assigned to the utterances must correspond with the interpretations the language user associates with them, i.e. they must be semantically and pragmatically relevant;
- the analyses should be in terms of what is traditional and familiar in descriptive linguistics so that they are readily accessible for linguists from various backgrounds.

The requirements that are listed above are fairly straightforward. In current corpus linguistic practice, however, where the parser is grammar-based and results from the automatic conversion of the grammar, the wish to meet all these requirements gives rise to some points of conflict of interest. Below we shall restrict ourselves to a discussion of those problems that are presently the most acute.

The first problem relates to the requirement of efficiency. The linguistic description of a natural language like English by means of a formal grammar yields a parser that is extremely complex. The parser being a product of, on the one hand, a substantial, linguistically motivated grammar and, on the other hand, a parsing algorithm that is not specifically geared to parsing complex ambiguous input, its efficiency is unpredictable, because so far there is too little knowledge and experience available about the combined effect of these two components. From our experiences so far it would seem, however, that a parser which is the result of the automatic conversion of a grammar written by a computationally naïve linguist is unlikely to be the most efficient parser that could be developed. To the linguist writing the grammar optimalization techniques that must be used in order to yield a more efficient parser are either unknown[16] or the motivation to use them is outweighed by the desire to let the grammar remain a linguistic object that is not obscured by matters of a computational nature. Here it must be pointed out that in our view the linguist in writing a grammar should have no need to concern himself with optimizing his description in the sense that it would yield a more efficient parser. While the linguist

[16] Also the computationally naïve linguist is frequently observed to step into the pitfalls of exponential complexity: the order in which certain rules or non-terminals occur, although they may -- from a linguistic point of view -- be irrelevant, is not always entirely free.

must yield an adequate linguistic description, it is the task of the computer scientist to provide the means to convert it into an optimally efficient parser.

The second problem relates to the linguistic felicity of the parser, which implies the requirements of full coverage, consistency, and semantic and pragmatic relevance. In writing our grammar we are faced with lacunae in our knowledge of (the description of) the language under investigation and problems raised by ill-defined notions in the (traditional) descriptions that are available. In order to deal with these problems it is necessary to write grammars that are more permissive than generative grammars that are (also) aimed at production. For example, the description of subject-verb concord that occurred in earlier versions of the TOSCA grammar was not incorporated in later versions after it had appeared that the rules governing grammatical subject-verb concord were frequently violated. In those instances where this was the case it appeared that some sort of proximity rule applied. However, the notion of proximity being an ill-defined one, it proved difficult to formalize it. Attempts at yielding a formal description of concord were abandoned. As a result, the grammar became more permissive in this respect than the earlier versions had been. Consequently, the parser analyzes sentences in which grammatical subject-verb concord is observed, as well as sentences in which concord based on the notion of proximity occurs. As a side-effect of the permissiveness of the grammar on this point, the parser will not fail to analyze a sentence which is unacceptable on account of the fact that there is no concord between the subject and the verb.

By allowing the parser to be more permissive in some respects the requirement of full coverage could be met: on those points where we fail to give a formal specification of the relations that hold between constituents -- whether on account of the lack of sufficient knowledge or due to problems of adequately formalizing what knowledge we do have -- we avoid the failure of analysis that would result from too restricted a description. A permissive parser should yield minimally the correct analysis/es for a given utterance. In the present context, where we restrict ourselves to the morpho-syntactic structure of the language, a correct analysis is one that emerges from the parser and can be associated with an interpretation that is contextually appropriate. Analyses that do not apply, i.e. analyses for which no interpretation can be found that is contextually adequate, must be discarded, while those that do are stored in the database.

The drawbacks of having a permissive parser are obvious: leaving details of the description unspecified, we permit the analysis of utterances that do not answer to the above notion of correctness and hence also allow a rather great degree of ambiguity. In either case interventions must compensate for the 'shortcomings'[17] of the parser. However, interventions are generally made on the basis of subjective judgments and errors readily occur. As a consequence, the consistency of analysis which in principle is warranted by the use of a non-probabilistic grammar may be affected.

After what was said in the previous paragraph we can now reformulate the requirements listed above in terms of three criteria that play a central role in the evaluation of a parser. The first criterion relates to the efficiency of the parser. An evaluation on this count should provide insight into the rate at which input can be parsed and hence the costs involved.[18] The second criterion concerns the nature of the input that can be parsed. Does the parser permit unrestricted input, or, if there are any restrictions with respect to the input, what is the nature of these restrictions? The third criterion relates to the output yielded by the parser. Questions that must be addressed are the following: (1) To what extent is the analysis consistent? Does the parser produce the same result for a given string at any one time? (2) What amount of input is successfully parsed, and in how many instances does the parser fail? (3) What is the quality of the analyses? This not only relates to the extent to which the analyses correspond with the interpretations that the language user associates with them, i.e. their semantic and pragmatic relevance, it also has to do with the amount of detail incorporated in the analyses. (4) Are the analyses readily accessible for linguists from various backgrounds?

The implementation of the above criteria which should provide us with a standard for assessing the grammar as an analysis tool, is partly easy, partly difficult. One problem lies in the fact that, where criteria lend

---

[17] Strictly speaking they are not shortcomings since they are the result of the decision to restrict the description to the morpho-syntactic structure of the language

[18] Note that the frequency and nature of the interventions must also be taken into consideration. We have refrained from listing it as a separate criterion since it possibly plays a role in the evaluation resulting from the application of each of the above criteria.

themselves to quantification, objective measures are yet to be established. An evaluation of the quality of analysis and also the accessibility of the analysis results is problematic since it does not seem possible to quantify these aspects. Therefore, an evaluation on these counts remains open to subjective judgments.

Quantification of criteria for evaluation

Finding measures by means of which the linguistic quality of the parser[19] can be quantified is less straightforward than one would be inclined to think. The same is true for measures that record the efficiency of the parser. Below we discuss various aspects that relate to the choice of measures in each of these instances.

Before we can concern ourselves with the question what particular measure is best in quantifying the linguistic quality of the parser, we need to consider to what end the parser is employed. This will give us an idea of what to count as successful. As we have said, in the context of current corpus linguistic practice our aim is to yield minimally the correct analysis/es[20] for a given string, while the input constitutes authentic language data, i.e. we aim to parse unrestricted input. A successful parse then is one for which the parser yields minimally the correct analysis/es. The extent to which the parser overgenerates, thus yielding additional analyses is of secondary importance. An unsuccessful parse results when, given an acceptable utterance, the parser does not yield a parse at all, or when the parser fails to yield an analysis that is contextually appropriate. Apart from successful and unsuccessful parses a third category of parses can be distinguished, namely the class of parses that remain inconclusive. This class comprises for instance parses that do not return within a given, reasonable amount of time. With a parser that shows exponential behaviour it does not make sense to insist on the analysis of each and every utterance at all costs, and a time-limit on each parse proves useful. However, time-limits can be

---

[19] The term 'linguistic quality of the parser' that is used here signifies both the nature of the input that the parser is capable of parsing as well as the quality of the resulting analyses.

[20] See above. A correct analysis is one that emerges from the parser and that can be associated with an interpretation that is contextually appropriate.

set arbitrarily and the failure of an analysis to return within a given amount of time is just that: it remains inconclusive whether the resultative parse would eventually be successful or not.

Having established what such notions as 'successful parse', 'unsuccessful parse' and 'inconclusive parse' signify, a measure must be found by means of which the parser's performance can be rated. In parsing natural language input different measures can be thought of. For example, the rate of success achieved can be expressed in terms of the number of words successfully parsed, the number of utterances successfully parsed, the longest utterance or even the largest number of levels of embedding parsed. The choice of what measure is employed in rating the linguistic quality of the parser is not merely a matter of preference, nor is it exclusively based on ideas regarding the usefulness of a particular measure. Some measures introduce a significant skewing factor in the process of evaluation. The success rate achieved in terms of the percentage of the total number of words as opposed to a rating in terms of the percentage of the total number of utterances may yield two widely divergent figures, depending on the length of the utterances in number of words. In fact, they can only be compared if the number of words is the same for all utterances involved, which will seldom be the case.

As to the quantification of the efficiency achieved by the parser we can distinguish between the overall efficiency of the parser on the one hand, and the efficiency of the parser in parsing particular structures. With respect to the former an account of the parsing times should suffice, provided we include full details about the machine the parser is run on. With respect to the latter, a record of the parsing times of individual utterances may give insight into the performance of the parser relative to the length of the input, or relative to the 'complexity' of the input. Here it must be observed that the notion of 'complexity' of the input is introduced somewhat tentatively in the awareness that, although this would probably give the most accurate idea of the parser's efficiency, this is a rather elusive concept.

5.4.2 The TOSCA grammar and parser

In this subsection some of the evaluative criteria that were discussed in the previous section are applied to the TOSCA grammar and the parser associated with it. Since an informal assessment of the grammar (both as a linguistic object and as an analysis tool) was already given in sec-

tion 5.3, we here focus on a more formal evaluation, which involves the quantification of various aspects of the grammar.

I. The grammar

It should be borne in mind that the grammar that was developed in the course of the TOSCA II project was from its inception intended for use in corpus analysis. As a consequence, design decisions were heavily influenced by the desire to develop a grammar that would yield a parser by means of which a syntactic database could be created, containing detailed information that would be readily accessible for linguists from various backgrounds. In this respect the choices that were made in selecting the formalism and the descriptive framework played an important role. The choice of Extended Affix Grammar was felicitous in the sense that the use of a generative grammar is familiar in linguistics. The descriptive framework that was used was based on the descriptive system put forward by Aarts and Aarts (1982), which in turn was an adaptation of the system found in Quirk et al. (1972). This system, in which a structure is assumed which is based on immediate constituency and the rank hierarchy, is subscribed to by a great many linguists who are concerned with the syntactic description of language.

The detail of description that is pursued can be measured in terms of the variety of labels denoting functions, categories and features. A complete list of these descriptive units can be found in Appendix F. In all, we distinguish as many as 78 functions, 64 categories -- 37 of which are lexical categories -- and 105 features. The richness of lexical categories and features, together with the fact that a great many words in English have multiple word class membership and/or can be associated with more than one feature set accounts for the amount of lexical distributional ambiguity that is encountered. The relative frequency of this type of ambiguity (as opposed to the syntactic ambiguity discussed below) is as follows: given the 1,000 most frequent words in English[21] we find that -- on average -- a word receives 1.89 analyses (in the range from one to six analyses per word), as far as its word class is concerned. If we also include the ambiguity caused by a word's features, the number of analyses per word increases to 3.96 (on average),

[21] That is, the 1,000 most frequent words according to the LOB frequency list compiled by Hofland and Johansson (1982).

while the range then appears to be 1 to 24 analyses per word.

In principle the grammar describes unrestricted English, i.e. there are no restrictions on the length of the utterances, nor are there any restrictions as to the structures or phenomena that are described. The adaptations that were made to the grammar in order to achieve a reduction in the amount of (syntactic) ambiguity yielded by the parser, or to prevent left-recursion, are -- in a sense -- restrictions. These, however, have little to do with the (un)restrictedness of the language described by the grammar; they are merely provisions that are made in the interest of the analysis process.

II. The parser

The evaluation of the parser that is presented here is based upon the analysis results that were obtained in parsing two samples from the TOSCA Corpus. Both samples contain slightly over 20,000 words.[22] One sample is fiction, the other non-fiction. The fiction sample (F) occurs in the text category 'crime fiction' and was taken from Michael Innes' *Carson's Conspiracy. A Sir John Appleby Mystery.* The non-fiction sample (N) was taken from an economic text by Dennis Swann, entitled *Competition and Consumer Protection.* We start by giving a more detailed characterization of the input.

The fiction sample consists of 1722 utterances in all, the non-fiction sample of 956 utterances. The mean utterance length (in terms of the number of orthographic words) in the fiction text is 11.56, while in the non-fiction text the mean length is 20.93. Here it must be observed that, while the mean length of utterances is commonly expressed in terms of the number of orthographic words they contain,[23] in the present context we prefer to measure the length of the utterances in terms of the number of tag units since the syntactic parser takes sequences of tags as input rather than sequences of orthographic words.[24]

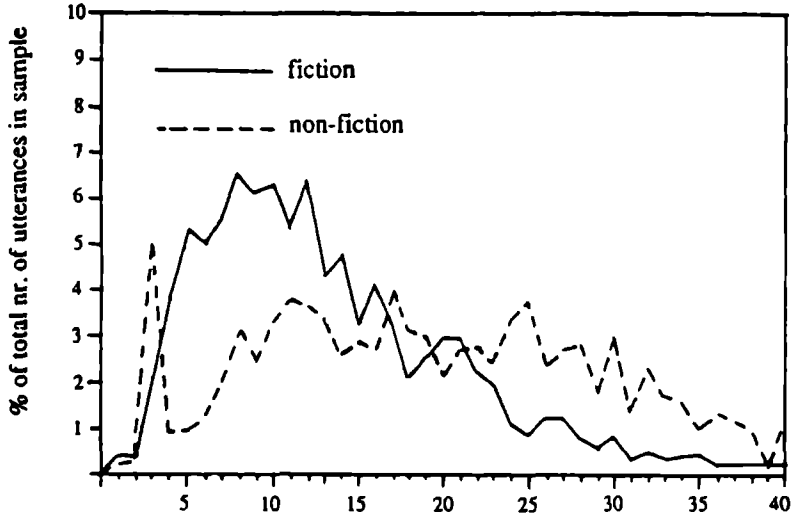
---

[22] 20,010 and 20,011 words respectively, to be exact.

[23] Cf. the study by Marckworth and Bell in Kucera and Francis (1967: 368-405), which was referred to in section 2.3.1.

Much more informative than the mean utterance length, however, is (the variation in) the length of the individual utterances. In Diagram 1 the samples are characterized in terms of the length of the utterances they contain. The relative distribution of the utterances by length (in terms of the number of tag units) shows that the fiction sample is far more internally homogeneous than the non-fiction sample.[25]

Diagram 1: Distribution of utterances by length



For the analysis of the samples a SUN 3-280 was used. In parsing the two samples the following results were obtained (cf. Table 1): of the

[24] The difference between an orthographic word and a tag unit is particularly relevant when a compound word is involved. For example, in the noun phrase 'a stimulus-response link up' the head is realized by a compound noun ('stimulus-response link up'). The tagging will associate a single tag with the compound word, whereas in terms of orthographic words we have three words. Note that the occurrence of compound words is not restricted to this one word class: there are compound adjectives (e.g. 'factor analytic', as in 'factor analytic studies'), verbs (e.g. 'give rise to'), adverbs (e.g. 'sort of', as in 'this sort of brings us'), etc.

[25] Full details on the distribution of utterances by length are given in Table A (Appendix G).

1722 utterances in the fiction sample 1517 (88.10%) received a correct analysis, for 78 (4.53%) the parser failed to yield an analysis, while the analysis of 127 utterances (7.38%) yielded an inconclusive result, i.e. the analysis of 106 utterances (6.16%) exceeded the time-limit that had been set at one hour CPU-time and for 21 utterances (1.22%) the output was too sizeable;[26] the analysis of the utterances contained in the non-fiction text proved to be much more problematic: of the 956 utterances 537 (56.17%) were correctly analyzed, for 117 utterances (12.24%) the parser failed to yield the correct analysis, while the analysis of the remaining 302 utterances (31.59%) yielded an inconclusive result, i.e. the analysis of 297 utterances was abandoned after the time-limit of one hour CPU-time had been exceeded and for 5 utterances (0.52%) the output was too sizeable.

Table 1: An overall quantification of the analysis results

sample	# utts	# successful	# inconclusive		# unsuccessful
			too sizeable	time up	
F	1722	1517 (88.10%)	21 (1.22%)	106 (6.16%)	78 (4.53%)
N	956	537 (56.17%)	5 (0.52%)	297 (31.07%)	117 (12.24%)

Table 2 gives a more detailed quantification of the analysis results. In the first column the length of the utterances is given in terms of the number of tag units. The second column lists the total number of utterances of a particular length, both for the fiction sample (F) and the non-fiction sample (N). The third column lists the number of utterances (of a particular length) that were successfully analyzed within the given time-limit of one hour CPU-time. Column four lists the number of utterances the analysis result of which yielded by the parser remained

[26] Here it should be observed that, although the parser succeeds in parsing the utterances, the total output (which is the product of the number of analyses times the sizes of the analyses) is too bulky for the working space we use (approximately 4 Mb) and the analyses must be discarded. It remains therefore inconclusive whether the result contains the correct analysis or not.

ıble 2: A more detailed quantification of the analysis results

inputlength	# utts		# successful		# inconclusive		# unsuccessful	
in # tag units	F	N	F	N	F	N	F	N
1	8	2	7	--	--	--	1	2
2	7	3	7	3	--	--	--	--
3	36	48	35	44	--	--	1	4
4	67	8	62	6	--	--	5	2
5	92	9	91	8	--	--	1	1
6	85	12	82	12	--	--	3	--
7	99	19	98	18	--	--	1	1
8	111	30	108	29	--	--	3	1
9	105	23	102	22	--	--	3	1
10	109	33	108	30	--	--	1	3
11	92	36	90	35	1	--	1	1
12	110	35	106	32	2	--	2	3
13	82	32	81	29	1	--	--	3
14	83	25	80	22	1	1	2	2
15	56	28	55	26	1	1	--	1
16	71	26	67	19	--	5	4	2
17	57	39	52	33	2	1	3	5
18	42	30	40	24	--	2	2	4
19	44	29	35	20	5	6	4	3
20	53	20	44	16	2	3	7	1
21	51	26	38	14	8	10	5	2
22	39	27	26	14	9	6	4	7
23	33	23	22	11	10	11	1	1
24	19	32	13	16	4	13	2	3
25	15	36	8	12	7	18	--	6
26	22	23	13	5	7	15	2	3
27	22	26	12	6	8	16	2	4
28	14	27	7	8	7	14	--	5
29	10	17	6	5	2	10	2	2
30	14	28	7	5	6	21	1	2
31	5	13	3	2	2	9	--	2
32	8	22	3	6	4	13	1	3
33	6	16	4	--	2	12	--	4
34	7	15	2	1	4	12	1	2
35	8	10	--	1	7	7	1	2
36	4	13	1	1	2	11	1	1
37	4	11	--	--	4	9	--	2
38	4	9	--	1	3	5	1	3
39	4	1	1	--	2	1	1	--
40	4	11	1	1	1	10	2	--
41-45	8	36	--	--	5	26	3	10
46-50	7	15	--	--	5	15	2	--
51-55	2	15	--	--	1	10	1	5
56-60	1	8	--	--	1	6	--	2
61-65	2	4	--	--	1	2	1	2
66-70	--	1	--	--	--	1	--	--
71-75	--	2	--	--	--	--	--	2
76-80	--	--	--	--	--	--	--	--
81-85	--	--	--	--	--	--	--	--
86-90	--	1	--	--	--	--	--	1
91-95	--	1	--	--	--	--	--	1

inconclusive. These are utterances for which the output of the analysis proved too sizeable and also utterances whose analysis exceeded the given time-limit. Finally, column five lists the number of utterances that the parser failed to analyze correctly.

As is apparent from Table 2 failure of analysis occurs with utterances of any length. Note also that failure of analysis appears to be a rather constant factor in the sense that there is no apparent correlation between the length of an utterance and the rate of success in analyzing it. Here it must be observed that where the parser fails to analyze a given utterance correctly this is either due to some error that occurred in the lexical-morphological tagging (including the syntactic pre-analysis) or to a lacuna in the grammar.[27]

In Table 3 a further specification can be found of the utterances the analysis of which yielded an inconclusive result.

The length of an utterance is significant when we consider the parsing time involved. As the figures in Table 3 show the time-limit is not exceeded for the analysis of utterances with a length up to 10 tag units. As the length of utterances increases the percentage of utterances the analysis of which must be abandoned after the time-limit of one hour CPU-time has been exceeded becomes increasingly larger.

In parsing the two samples discussed here the time-limit was set at one hour CPU-time. Earlier we already observed that time-limits can be set arbitrarily. Therefore the question that must be raised is: to what extent is the given time-limit arbitrary? or, put differently, what would be the effect if the time-limit were extended? An evaluation of the parsing times that were achieved by the parser shows that the time-limit of one hour is not all that arbitrary: in the case of the fiction sample the maximum time that was needed to yield a conclusive result was 3000 CPU-seconds; in the case of the non-fiction sample parsing times that exceeded 1800 CPU-seconds did not result in any successful analyses. The parser typically shows exponential behaviour: an increase in the length of the input causes the amount of time that is needed to parse the input to increase exponentially, i.e. the length of the input is an exponential factor in the (time)function which characterizes the parser's efficiency. Diagram 2 shows the relative gain in number of utterances analyzed set against the cost in time.

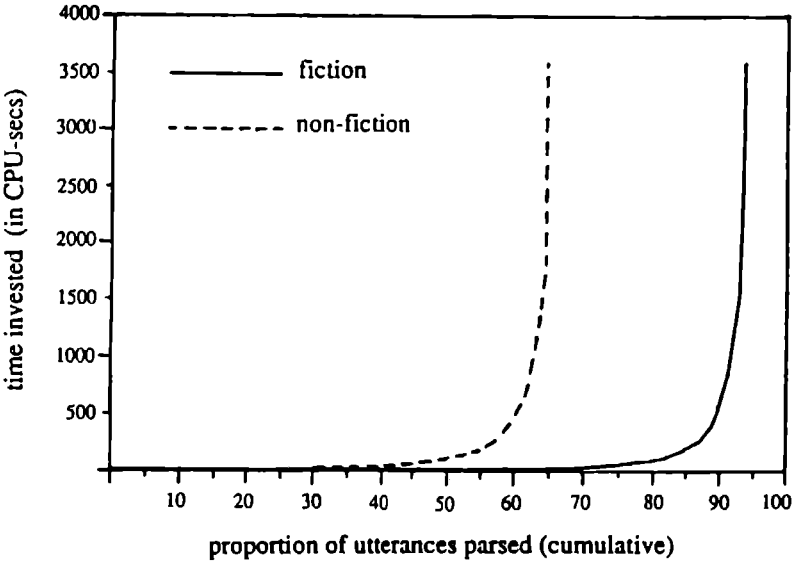
---

[27] Of course, where the parser fails to yield an analysis for an utterance that must be considered unacceptable, this should not be ascribed to a shortcoming of the grammar.

ıble 3: A further specification of inconclusive parses

inputlength	# utts		# too sizeable		# time up	
in # tag units	F	N	F	N	F	N
1	--	--	--	--	--	--
2	--	--	--	--	--	--
3	--	--	--	--	--	--
4	--	--	--	--	--	--
5	--	--	--	--	--	--
6	--	--	--	--	--	--
7	--	--	--	--	--	--
8	--	--	--	--	--	--
9	--	--	--	--	--	--
10	--	--	--	--	--	--
11	1	--	--	--	1	--
12	2	--	--	--	2	--
13	1	--	--	--	1	--
14	1	1	--	--	1	1
15	1	1	1	--	--	1
16	--	5	--	--	--	5
17	2	1	1	--	1	1
18	--	2	--	--	--	2
19	5	6	--	--	5	6
20	2	3	--	--	2	3
21	8	10	2	--	6	10
22	9	6	4	1	5	5
23	10	11	2	--	8	11
24	4	13	1	1	3	12
25	7	18	1	--	6	18
26	7	15	1	--	6	15
27	8	16	2	--	6	16
28	7	14	1	2	6	12
29	2	10	--	--	2	10
30	6	21	--	1	6	20
31	2	9	--	--	2	9
32	4	13	--	--	4	13
33	2	12	--	--	2	12
34	4	12	--	--	4	12
35	7	7	1	--	6	7
36	2	11	1	--	1	11
37	4	9	--	--	4	9
38	3	5	1	--	2	5
39	2	1	2	--	--	1
40	1	10	--	--	1	10
41-45	5	26	--	--	5	26
46-50	5	15	--	--	5	15
51-55	1	10	--	--	1	10
56-60	1	6	--	--	1	6
61-65	1	2	--	--	1	2
66-70	--	1	--	--	--	1
71-75	--	--	--	--	--	--
76-80	--	--	--	--	--	--
81-85	--	--	--	--	--	--
86-90	--	--	--	--	--	--
91-95	--	--	--	--	--	--

Diagram 2: Cumulative proportion of conclusive results in relation to analysis time invested



While at present the analysis of a relatively large percentage of utterances remains inconclusive and must await further optimalization of the parser, it must also be observed that 37.03% of the utterances in the fiction sample and 16.95% of the utterances in the non-fiction sample were parsed within five seconds. 70.55% (F) / 41.35% (N) of the utterances yielded a parse within 30 seconds; 75.42% (F) / 46.75% (N) of the utterances was parsed within one minute and 80.10% (F) / 49.76% (N) within 90 seconds. Full details are given in Table 4.

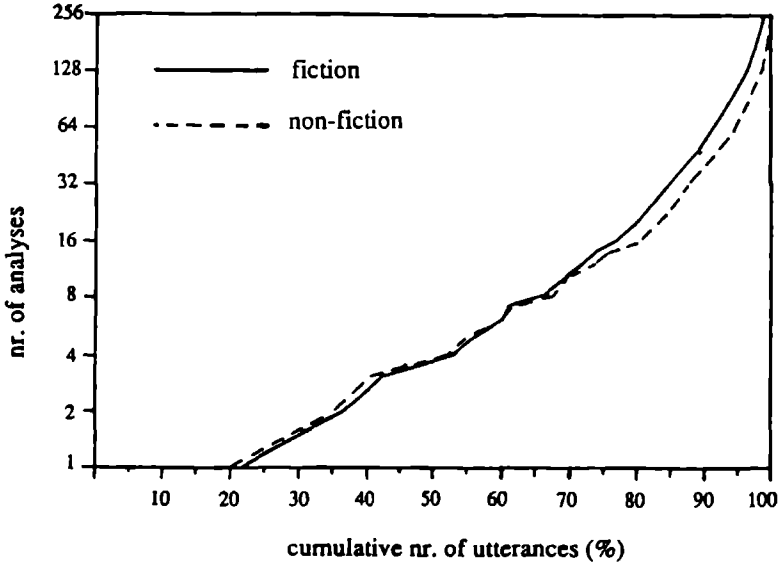
The amount of syntactic ambiguity that was yielded by the parser in analyzing the fiction and the non-fiction sample is represented in Diagram 3.

In terms of the amount of syntactic ambiguity the two samples were very much alike. About 20% of the utterances (21.95% in the case of the fiction sample, 20.11% in the case of the non-fiction sample) received a single analysis. Although this leaves some 80% of the utterances which yielded multiple analyses, extreme cases of syntactic ambiguity remained rare.[28]

ıble 4: Parsing times achieved

time in	# utts		% utts		cum. # utts		cum. % utts	
CPU-secs.	F	N	F	N	F	N	F	N
≤ 5	601	141	37.03	16.95	601	141	37.03	16.95
10	303	92	18.67	11.06	904	233	55.70	28.00
15	114	50	7.02	6.01	1018	283	62.72	34.01
20	59	27	3.64	3.25	1077	310	66.36	37.26
25	37	16	2.28	1.92	1114	326	68.34	39.18
30	31	18	1.91	2.16	1145	344	70.55	41.35
35	16	14	0.99	1.68	1161	358	71.53	43.03
40	22	8	1.36	0.96	1183	366	72.89	43.99
45	10	5	0.62	0.60	1193	371	73.51	44.59
50	4	8	0.25	0.96	1197	379	73.75	45.55
55	18	6	1.11	0.72	1215	385	74.86	46.27
60	9	4	0.55	0.48	1224	389	75.42	46.75
65	16	9	0.99	1.08	1240	398	76.40	47.84
70	9	6	0.55	0.72	1249	404	76.96	48.56
75	15	3	0.92	0.36	1264	407	77.88	48.92
80	15	2	0.92	0.24	1279	409	78.80	49.16
85	10	4	0.62	0.48	1289	413	79.42	49.64
90	11	1	0.68	0.12	1300	414	80.10	49.76
100	9	7	0.55	0.84	1309	421	80.65	50.60
110	13	6	0.80	0.72	1322	427	81.45	51.32
120	13	5	0.80	0.60	1335	432	82.26	51.92
130	10	3	0.62	0.36	1345	435	82.87	52.28
140	2	5	0.12	0.60	1347	440	82.99	52.88
150	7	8	0.43	0.96	1354	448	83.43	53.85
160	5	2	0.31	0.24	1359	450	83.73	54.09
170	4	2	0.25	0.24	1363	452	83.98	54.33
180	9	4	0.55	0.48	1372	456	84.53	54.81
210	20	9	1.23	1.08	1392	465	85.77	55.89
240	18	6	1.11	0.72	1410	471	86.88	56.61
270	12	7	0.74	0.84	1422	478	87.62	57.45
300	10	2	0.62	0.24	1432	480	88.23	57.69
360	9	11	0.55	1.32	1441	491	88.79	59.01
420	9	6	0.55	0.72	1450	497	89.34	59.74
480	9	5	0.55	0.60	1459	502	89.90	60.34
540	2	7	0.12	0.84	1461	509	90.02	61.18
600	7	5	0.43	0.60	1468	514	90.45	61.78
660	4	2	0.25	0.24	1472	516	90.70	62.02
720	3	--	0.18	0.00	1475	516	90.88	62.02
780	2	2	0.12	0.24	1477	518	91.00	62.26
840	4	--	0.25	0.00	1481	518	91.25	62.26
900	2	2	0.12	0.24	1483	520	91.37	62.50
1200	11	8	0.68	0.96	1494	528	92.05	63.46
1500	11	2	0.68	0.24	1505	530	92.73	63.70
1800	5	5	0.31	0.60	1510	535	93.04	64.30
2100	1	--	0.06	0.00	1511	535	93.10	64.30
2400	2	--	0.12	0.00	1513	535	93.22	64.30
2700	1	--	0.06	0.00	1514	535	93.28	64.30
3000	2	--	0.12	0.00	1516	535	93.41	64.30
3600	--	--	0.00	0.00	1516	535	93.41	64.30
time up	106	297	6.53	35.70	1623	832	100.00	100.00

Diagram 3: Syntactic ambiguity yielded by the parser



One aspect that has not entered into the discussion so far is the fact that the input to the syntactic parser was virtually unambiguous as far as the lexical-morphological tagging was concerned, while as a result of the syntactic pre-analysis a reduction of the amount of syntactic ambiguity was achieved. As a side-effect of the interventions that were made the efficiency of the parser increased. The reader will recall that on several occasions the role of interventions has been commented upon. It has been argued that, whenever possible, they should be avoided since they are liable to affect the consistency of analysis and are apt to introduce errors. The syntactic pre-analysis that was carried out in the case of the two samples discussed here can be quantified as follows: in the case of the fiction sample 42.28% of the utterances remained unmarked, while in the non-fiction sample 19.56% of the utterances did not receive a marker. In 29.27% (F) / 20.82% (N) of the utterances one constituent was marked, and in 15.16% (F) / 17.68% (N) two constituents were marked.[29] The apparent difference in the degree of marking between the

[28] See also Table B (Appendix G).

[29] Appendix H lists the syntactic markers that were employed in the syntactic pre-analysis. In

fiction and the non-fiction sample relates to the length and also the structural complexity of the utterances in the respective samples.

5.5 What still has to be done

From our discussion of the analysis results above and the evaluation of some of the problems that were encountered, it is apparent that in spite of the fact that considerable progress has been made over the past few years, there is still a lot that needs to be done. For the time being the creation of sufficiently large and reliable databases for linguistic research must continue to be our main concern. The availability of increasingly more analyzed data and the experiences and insights acquired in the process of producing them will give further direction to future research.

Even to date it is still true that the availability of corpora that have received a detailed (syntactic) analysis that fully conforms to the reader's / listener's semantic and pragmatic interpretation of the analyzed utterances leaves much to be desired. So far only the rather small (130,000 words) Nijmegen Corpus that was analyzed in the pilot project (CCPP) which preceded the TOSCA projects, has received such a detailed analysis. At the present time the Nijmegen TOSCA Corpus is in the process of being analyzed. Upon completion the analyzed corpus will grant access to a body of text comprising some 1.5 million words in all. Exploration of this material is expected to yield a wealth of information, including information that before could not be obtained at all or only on a small scale. Structures that have received very little attention in the handbooks on English grammar can be studied more extensively, while such matters as the frequency and distribution of occurrence of various syntactic structures can be investigated. Moreover, the composition of the corpus should make it possible to investigate aspects of linguistic variation. Texts within one and the same text category can be compared with each other, and contrasted with texts from other text categories. Not until sufficient information has been gathered, however, regarding various structures and their variants, and

---

this appendix also information is provided on the number of markers that were introduced per utterance. The marker that was most frequently used is that marking the end of a noun phrase postmodifier.

the linguistic and extra-linguistic factors that account for the variation encountered, can the development of a varieties grammar be held to be a feasible undertaking. Meanwhile, we must aim at the development of a descriptive framework that lends itself to the description of linguistic variation, and investigate and experiment with formalisms that will make it possible to formalize this type of description.

Apart from the exploration of the analyzed material, future research should comprise an extension of the scope of the grammar and of descriptive theory. So far the analysis of corpora has been restricted to the morpho-syntactic analysis of individual utterances against the background of a descriptive framework that is both traditional and familiar to the potential future users of the corpus. The formalized descriptions contained in the grammars should be extended so as to include aspects of a semantic and pragmatic nature. It should also be attempted to formalize the description of text structure and cohesion. Whether or not such a description can be integrated into our present grammars that describe the structure of individual utterances, is a question that cannot now be answered.

Whereas the full-scale exploration of the corpora and the extension of the grammar belong to the tasks that will be undertaken in due time, we find that at present other tasks are more pressing. These require the joint effort of computer scientists and linguists. While the latter should concern themselves with the exploration of material already analyzed in search of linguistic facts that will contribute to reducing the amount of ambiguity generated by the grammar, it is the task of the computer scientists to develop parser generators that yield better parsers. Since the results of these developments will not be available overnight, other, less ambitious, solutions must be considered. For example, it may be worthwhile to adapt the grammar on points where structural vagueness must be held responsible for the recurrent ambiguity found with particular structures. A set of rules which leaves this ambiguity unspecified, to be resolved at a later stage, would contribute to reducing the complexity of the parser. Another direction in which a solution may be sought is that of making available tools that can be helpful in optimizing the intervention process.

The above suggestions for the direction future research should take derive from the experiences gained and the insights acquired in the research projects reported on in this book. In the account that was pre-

sented of the approach taken in the TOSCA projects both achievements and problems have been highlighted. It was pointed out that, unlike other undertakings in the processing of authentic (natural) language data, the Nijmegen approach has been less concerned with pursuing the immediate incorporation of its results in some application or other. Instead a serious attempt has been made at parsing truly unrestricted input while adhering to a high standard of linguistic quality in its analyses. Despite the fact that a great deal of work is yet to be done, it is our belief that the methodological principles that underlie the corpus linguistic approach pursued in Nijmegen has provided this approach to natural language processing with a firm foundation, which is methodologically sound as well as practically and linguistically feasible. We may therefore expect future research to provide more insight into the nature of language use, the characterization of language varieties and hence achieve a significant contribution to a linguistic theory about language use.

References cited

Aarts, F.G.A.M. (1976): "The description of linguistic variation in English: From Firth till the present day", in: *English Studies* 57 (1976): 239-251.

Aarts, F.G.A.M. (1984): "Linguistic variation in English: Idealization, variety and linguistic items", in: *English Studies* 65 (1984): 59-76.

Aarts, F. and J. Aarts (1982): *English Syntactic Structures. Functions and Categories in Sentence Analysis.* Oxford: Pergamon Press Ltd.

Aarts, J. (1980): *Taalkunde en Hedendaags Engels.* Utrecht: Bohn, Scheltema en Holkema.

Aarts, J. (1984a): "The LDB: A linguistic database, in: *ICAME News* 8 (1984): 25-30.

Aarts, J. (1984b): "The description of the English language", in: Mackenzie and Wekker (eds.) (1984): 13-32.

Aarts, J. (1990): "Corpus linguistics: An appraisal", in: Choueka (ed.) (1990): 13-28.

Aarts, J. (1991): "Intuition-based and observation-based grammars", to appear in: Aijmer and Altenberg (eds.) (1991).

Aarts, J. and Th. van den Heuvel (1982): "Grammars and intuitions in corpus linguistics", in: Johansson (ed.) (1982): 66-84.

Aarts, J. and Th. van den Heuvel (1983): "Corpus-based syntax studies", in: *Gramma* 7 (1983): 153-173.

Aarts, J. and Th. van den Heuvel (1984): "Linguistic and computational aspects of corpus research", in: Aarts and Meijs (eds.) (1984): 83-94.

Aarts, J. and Th. van den Heuvel (1985): "Computational tools for the syntactic analysis of corpora", in: *Linguistics* 23 (1985): 303-335.

Aarts, J. and W. Meijs (eds.) (1984): *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research.* Amsterdam: Rodopi.

Aarts, J. and W. Meijs (eds.) (1986): *Corpus Linguistics II. New Studies in the Analysis and Exploitation of Computer Corpora.* Amsterdam: Rodopi.

Aarts, J. and W. Meijs (1989): "Corpustaalkunde", in: *Spektator* 18-1: 6-23.

Aarts, J. and W. Meijs (eds.) (1990): *Theory and Practice in Corpus Linguistics.* Amsterdam: Rodopi.

Aarts, J. and N. Oostdijk (1988): "Corpus-related research at Nijmegen University", in: Kytö, Ihalainen and Rissanen (eds.) (1988): 1-14.

Abercrombie, D. (1955): "English Accents", in: *The Speech Teacher* 4: 10-18.

Aijmer, K. and B. Altenberg (eds.) (1991): *English Corpus Linguistics: Studies in honour of Jan Svartvik.* London: Longman.

Allen, J.P.B. and S. Pit Corder (eds.) (1973): *The Edinburgh Course in Applied Linguistics.* Vol. I: Readings for Applied Linguistics. Oxford: OUP.

Atwell, E. (1987): "Constituent-likelihood grammar", in: Garside, Leech and Sampson (eds.) (1987): 57-65.

Atwell, E., G. Leech and R. Garside (1984): "Analysis of the LOB Corpus: Progress and prospects", in: Aarts and Meijs (eds.) (1984): 41-52.

Bailey, C.N. and R.W. Shuy (eds.) (1974, 1973): *New Ways of Analyzing Variation in English.* Washington: Georgetown University Press.

Bergenholtz, H. and B. Schaeder (eds.) (1979): *Empirische Textwissenschaft. Aufbau und Auswertung von Textcorpora.* Koenigstein: Scriptor.

Berwick, R. (1987): "Transformational Grammar", in: Shapiro, S. (ed.) (1987): 353-361.

Biber, D. (1985): "Investigating macroscopic textual variation through multifeature/multidimensional analyses", in: *Linguistics* 23 (1985): 337-360.

Biber, D. and E. Finegan (1986): "An initial typology of English text types", in: Aarts and Meijs (eds.) (1986): 19-46.

Bresnan, J. (1974): "The position of certain clause-particles in phrase structure", in: *Linguistic Inquiry* 5 (1974): 614-619.

Briscoe, T. (1990): "English noun phrases are regular: a reply to Professor Sampson", in: Aarts and Meijs (eds.) (1990): 45-60.

Carroll, J.B. (1960): "Vectors in prose style", in: Sebeok (ed.) (1960): 283-292.

Catford, J.C. (1962): *Language Varieties.* A paper given to the Linguistic Association of Great Britain.

Catford, J.C. (1965): *A Linguistic Theory of Translation.* London: Oxford University Press.

Chambers, J.K. and P. Trudgill (1980): *Dialectology.* Cambridge: Cambridge University Press.

Charniak, E. (1981): "A parser with something for everyone", in: King, M. (ed.) (1981): 117-150.

Chomsky, N. (1957): *Syntactic Structures.* The Hague: Mouton.

Choueka, Y. (ed.) (1990): *Computers in Literary and Linguistic Research * Literary and Linguistic Computing 1988.* Proceedings of the Fifteenth International Conference, Jerusalem 5-9 June 1988. Paris-Geneva: Champion-Slatkine.

Crystal, D. (1969): *Prosodic Systems and Intonation in English.* Cambridge: Cambridge University Press

Crystal, D. and D. Davy (1969): *Investigating English Style.* London: Longmans.

Crystal, D. and D. Davy (1973): "Stylistic analysis", in: Allen and Pit Corder (eds.) (1973): 69-90.

Crystal, D. and R. Quirk (1964): *Systems of Prosodic and Paralinguistic Features in English.* Janua Linguarum, Series Minor 39. The Hague: Mouton.

Ditters, E. (1987): "Progress report ASCAMSA", in: *Processing Arabic* 2 (1987): 40-51.

Eeg-Olofsson, M. and J. Svartvik (1984): "Four-level tagging of spoken English", in: Aarts and Meijs (eds.) (1984): 53-64.

Ellegard, A. (1978): *The Syntactic Structure of English Texts.* A computer-based study on four kinds of text in the Brown University Corpus. Gothenburg: Acta Universitatis Gothoburgensis.

Ellis, J. and J.N. Ure (1969): "Language varieties: Register", in: Meetham and Hudson (eds.) (1969): 251-259.

Evans, R. (1985): "ProGram -- a development tool for GPSG grammars", in: *Linguistics* 23 (1985): 213-243.

Francis, W.N. (1964): *Manual of Information to Accompany a Standard Sample of Present-Day Edited American English for Use with Digital Computers.* Providence, Rhode Island: Dept. of Linguistics, Brown University.

Francis, W.N. (1979): "Problems of assembling and computerizing large corpora", in: Bergenholtz and Schaeder (eds.) (1979): 110-123.

Francis, W.N. (1980): "A tagged corpus -- problems and prospects", in: Greenbaum, Leech and Svartvik (eds.) (1980): 192-209.

Francis, W.N. and H. Kucera (1983): *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston.

Friedman, J. (1969): "A computer system for transformational grammar", in: *Communications of the ACM* 12 (1969): 341-348.

Friedman, J. (1978): "Computational and theoretical studies in Montague grammar at the University of Michigan", in: *SISTM Quarterly* 1 (1978): 62-66.

Garside, R., G. Leech and G. Sampson (eds.) (1987): *The Computational Analysis of English. A corpus-based approach*. London: Longman.

Gazdar, G. (1985): "Computational tools for doing linguistics", in: *Linguistics* 23 (1985): 185-187.

Gazdar, G., E. Klein, G. Pullum and I. Sag (1985): *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Gorsuch, R.L. (1983): *Factor Analysis*. Hillsdale, N.J.: Erlbaum.

Greenbaum, S. (1984): "Corpus analysis and elicitation tests", in: Aarts and Meijs (eds.) (1984): 193-201.

Greenbaum, S. (1988): *Good English and the Grammarian*. London: Longman.

Greenbaum, S., G. Leech and J. Svartvik (eds.) (1980): *Studies in English Linguistics for Randolph Quirk*. London: Longman

Gregory, M. (1967): "Aspects of varieties differentiation", in: *Journal of Linguistics*, 3 (1967): 177-198.

Haan, P. de (1984): "Relative clauses compared", in: *ICAME News* 8 (1984): 47-59.

Haan, P. de (1989): *Postmodifying Clauses in the English Noun Phrase. A Corpus-Based Study*. Amsterdam: Rodopi.

Haan, P. de and R. van Hout (1986): "Statistics and corpus analysis", in: Aarts and Meijs (eds.) (1986): 79-98.

Hallebeek, J. (1990): *Een grammatica voor automatische analyse van het Spaans.* Nijmegen: University of Nijmegen (Doctoral thesis).

Halteren, H. van and Th. van den Heuvel (1990): *Linguistic Exploitation of Syntactic Databases. The Use of the Nijmegen Linguistic DataBase Program.* Amsterdam: Rodopi.

Halteren, H. van and N. Oostdijk (1988): "Using an analyzed corpus as a linguistic database", in: Roper (ed.) (1988): 171-181.

Halvorsen, P. (1988): "Computer applications of linguistic theory", in: Newmeyer (ed.) (1988): 198-219.

Harris, Z. (1951): *Methods in Structural Linguistics.* Chicago: University of Chicago Press.

Heuvel, Th. van den (1987): "Interaction in syntactic corpus analysis", in: Meijs (ed.) (1987): 235-252.

Heuvel, Th. van den (1988): "TOSCA: An aid for building syntactic databases", in: *Literary and Linguistic Computing* 3 (1988): 147-151.

Hofland, K. and S. Johansson (1982): *Word Frequencies in British and American English.* Bergen: Norwegian Computer Centre for the Humanities.

Hudson, R. (1980): *Sociolinguistics.* Cambridge: Cambridge University Press.

Hudson, R. (1981): "Some issues on which linguists can agree", in: *Journal of Linguistics* 17 (1981): 333-343.

Jespersen, O. (1909-49): *A Modern English Grammar on Historical Principles.* Copenhagen: Munksgaard.

Johansson, S. with G. Leech and H. Goodluck (1978): *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* Oslo: Dept. of English, University of Oslo.

Johansson, S. (1979): "The use of a corpus in register analysis: The case of learned and scientific English", in: Bergenholtz and Schaeder (eds.) (1979): 281-293.

Johansson, S. (ed.) (1982): *Computer Corpora in English Language Research.* Bergen: Norwegian Computer Centre for the Humanities.

Johansson, S. in collaboration with E. Atwell, R. Garside and G. Leech (1986): *The Tagged LOB Corpus. User's Manual.* Bergen: Norwegian Computing Centre for the Humanities.

Kaye, G. (1987): *Survey of English Usage. Computerization.* Winchester: IBM Scientific Centre.

Keulen, F. (1986): "The Dutch Computer Corpus Pilot Project. Some experiences with a semi-automatic analysis of contemporary English", in: Aarts and Meijs (eds.) (1986): 127-161.

King, M. (1983): *Parsing Natural Language.* London: Academic Press.

Koster, C. (1971): "Affix Grammars", in: Peck (ed.) (1971): 95-109.

Koster, C. (1991): *Affix Grammars for Natural Languages.* Nijmegen: University of Nijmegen.

Kruisinga, E. (1909-32): *A Handbook of Present-Day English.* Groningen.

Kucera, H. and W. Nelson Francis (1967): *Computational Analysis of Present-Day American English.* Providence, Rhode Island: Brown University Press.

Kühling, P. (1978): *Affix-Grammatiken zur Beschreibung von Programmiersprachen.* Berlin: Technische Universität.

Kytö, M., O. Ihalainen and M. Rissanen (eds.) (1988): *Corpus Linguistics Hard and Soft.* Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi.

Labov, W. (1979, 1972): *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.

Langendoen, T. (1969): Review article on Bazell et al.: *In memory of J.R. Firth* (Longmans, 1966), in: *Foundations of Language* 5 (1969): 391-408.

Leech, G. (1987): "The computational analysis of English. General Introduction", in: Garside et al. (1987): 1-15.

Leech, G., R. Garside and E. Atwell (1983): "The automatic grammatical tagging of the LOB Corpus", in: *ICAME News* 7 (1983): 13-33.

Mackenzie, J. and H. Wekker (eds.) (1984): *English Language Research.* Amsterdam: Free University Press.

Marckworth, M.L. and L.M. Bell (1967): "Sentence-length distribution in the Corpus", in: Kucera and Francis (eds.) (1967): 368-405.

Marcus, M. (1980): *A Theory of Syntactic Recognition for Natural Language.* Cambridge, Mass.: MIT Press.

Meetham, A.R. and R.A. Hudson (eds.) (1969): *Encyclopaedia of Linguistics, Information and Control.* Oxford: Pergamon Press.

Meijer, H. (1986): *Programmar: A translator generator.* Nijmegen: University of Nijmegen (Doctoral thesis).

Meijs, W. (ed.) (1987): *Corpus Linguistics and Beyond.* Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi.

Neijt, A. (1979): *Gapping. A Contribution to Sentence Grammar.* Dordrecht: Foris Publications.

Newmeyer, F. (1983): *Grammatical Theory. Its Limits and Its Possibilities.* Chicago: The University of Chicago Press.

Newmeyer, F. (ed.) (1988): *Linguistics: The Cambridge Survey.* Vol. II Linguistic Theory: Extensions and Implications. Cambridge: Cambridge University Press.

Oostdijk, N. (1983): *Extended Affix Grammar and English Syntax: A Formal Description of the English Noun Phrase.* Unpublished MA thesis. Nijmegen: Dept. of English.

Oostdijk, N. (1984): "An extended affix grammar for the English noun phrase", in: Aarts and Meijs (eds.) (1984): 95-122.

Oostdijk, N. (1986): "Coordination and gapping in corpus analysis", in: Aarts and Meijs (eds.) (1986): 177-202.

Oostdijk, N. (1988a): "A corpus linguistic approach to linguistic variation", in: *Literary and Linguistic Computing* 3 (1988): 12-25.

Oostdijk, N. (1988b): "A corpus for studying linguistic variation", in: *ICAME Journal* 12 (1988): 3-14.

Oostdijk, N. (1988c): "Corpustaalkunde in perspectief", in: *TTT* 8 (1988): 209-224.

Oostdijk, N. (1989): *TOSCA Corpus -- Manual.* (Preliminary version) Nijmegen: University of Nijmegen.

Oostdijk, N. (1990a): "Ambiguity in syntactic corpus analysis", in: Choueka (ed.) (1990): 315-333.

Oostdijk, N. (1990b): "The language of dialogue in fiction", in: *Literary and Linguistic Computing* 5 (1990): 235-241.

Oostdijk, N. (to appear): *TOSCA Syntax -- Manual.* Nijmegen: University of Nijmegen.

Peck, J. (ed.) (1971): *ALGOL68 Implementation.* Amsterdam: North Holland.

Phillips, J. and H. Thompson (1985): "GPSGP -- a parser for generalized phrase structure grammars", in: *Linguistics* 23 (1985): 245-261.

Poutsma, H. (1904-26): *A Grammar of Late Modern English.* Groningen: P. Noordhoff.

Quirk, R. (1968): *Essays on the English Language. Medieval and Modern.* London: Longman.

Quirk, R. (1974): *The Linguist and the English Language.* London: Edward Arnold.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1972): *A Grammar of Contemporary English.* London: Longman.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985): *A Comprehensive Grammar of the English Language.* London: Longman.

Quirk, R. and J. Svartvik (1979): "A corpus of Modern English", in: Bergenholtz and Schaeder (eds.) (1979): 204-218.

Renouf, A. (1984): "Corpus development at Birmingham University", in: Aarts and Meijs (eds.) (1984): 3-39.

Reenen-Stein, K. van, P. van Reenen, and A. Dees (eds.) (1988): *Corpusgebaseerde Woordanalyse.* Jaarboek 1987-1988. Amsterdam: Vrije Universiteit.

Roper, J. (ed.) (1988): *Computers in Linguistic and Literary Computing.* Proceedings of the Thirteenth ALLC Conference, University of East Anglia (Norwich) 1-4 April 1986. Paris-Geneva: Champion-Slatkine.

Sager, N. (1981): *Natural Language Information Processing.* A computer grammar of English and its applications. Reading, Mass.: Addison-Wesley Publishing Company.

Sampson, G. (1987a): "Probabilistic models of analysis", in: Garside et al. (1987): 16-29.

Sampson, G. (1987b): "Evidence against the 'grammatical'/'ungrammatical' distinction", in: Meijs (ed.) (1987): 219-226.

Sebeok, T.A. (ed.) (1960): *Style in Language.* Cambridge, Mass.: The M.I.T. Press.

Shapiro, S. (ed.) (1987): *Encyclopaedia of Artificial Intelligence.* New York: Wiley-Interscience Publication.

Shieber, S. (1985): "Criteria for designing computer facilities for linguistic analysis", in: *Linguistics* 23 (1985): 189-211.

Sinclair, J. (1982): "Reflections on computer corpora in English language research", in: Johansson (ed.) (1982): 1-6.

Svartvik, J. (ed.) (1990): *The London-Lund Corpus of Spoken English. Description and Research.* Lund: Lund University Press.

Svartvik, J. and R. Quirk (eds.) (1980): *A Corpus of English Conversation.* Lund Studies in English 56. Lund: Gleerup.

Svartvik, J., M. Eeg-Oloffson, O. Forsheden, B. Orestrom, and C. Thavenius (1982): *Survey of Spoken English. Report on Research 1975-81.* Lund Studies in English 1963. Lund: Gleerup.

Taylor, L., C. Grover and E. Briscoe (1989): "The syntactic regularity of English noun phrases", in: *Proceedings of the 4th European Chapter of the Association for Computational Linguistics held at Manchester*: 256-63.

Watt, D. (1974): *Analysis-Oriented Two-Level Grammars.* Glasgow: University of Glasgow (Ph.D. Thesis).

Winograd, T. (1983): *Language as a Cognitive Process.* Vol. 1 Syntax. Reading, Mass.: Addison-Wesley Publishing Company. Ch.7: Computer Systems for Natural Language Parsing.

Zwol, A.J. van (1990): *AGFLs Revisited. A parser system with integrated lexicon.* Nijmegen: University of Nijmegen.

Appendix A

The final design of the Survey of Educated English Usage

The design of the corpus comprises 'texts' of 5,000 running words each (cf. Table 1 in Quirk and Svartvik, 1979: 207).

I Material with origin in writing (100 texts)

A Printed (46)

Learned arts	6
Learned sciences	7
Instructional	6
Press: general news	4
specific reporting	4
Administrative & official material	4
Legal and statutory material	3
Persuasive writing	5
Prose fiction	7

B Non-printed (36)

Continuous writing:	imaginative	5
	informative	6
Letters: social	intimate	6
	equal	3
	distant	4
Letters: non-social	equal	4
	distant	4
Personal journals (diaries)		4

C As Spoken (18)

Drama	4
Formal scripted oration	3
Broadcast news	3
Talks: informative	4
imaginative	2
Stories	2

II Material with origin in speech (100 texts)

A Monologue (24)		B Dialogue (76)		
Prepared (but unscripted)		Conv. surreptitious	intimate	24
oration	6		distant	8
Spontaneous:		Conv. non-surrept.	intimate	22
oration	10		distant	6
commentary sport	4	Conv. telephone	intimate	10
non-sport	4		distant	6

Appendix B

Contents of the London-Lund Corpus

Face-to-face conversation	(46)	
Subgroup A	(34)	
dialogue		+
face-to-face		+
private		+
surreptitious		+
radio		-
Subgroup B	(12)	
dialogue		+
face-to-face		+
private		+
surreptitious		-
radio		-
Telephone conversation	(10)	
Subgroup C	(10)	
dialogue		+
face-to-face		-
private		+
surreptitious		+
radio		-
Discussion, interview, debate	(12)	
Subgroup D	(12)	
dialogue		+
face-to-face		+
private		-
surreptitious		-
radio		+

Public, unprepared commentary, demonstration, oration	(12)	
Subgroup E	(3)	
dialogue		+
face-to-face		+
private		-
surreptitious		-
radio		-
Subgroup F	(2)	
dialogue		-
face-to-face		+
private		-
surreptitious		-
radio		+
Subgroup G	(7)	
dialogue		-
face-to-face		-
private		-
surreptitious		-
radio		+
Public, prepared oration	(7)	
Subgroup H	(7)	
dialogue		-
face-to-face		+
private		-
surreptitious		-
radio		-

Appendix C

Contents of each major text category in the Brown Corpus

I. Informative Prose (374 samples)

Category A Press: reportage

Political	Daily	10	Weekly	4	Total	14
Sports		5		2		7
Society		3		0		3
Spot news		7		2		9
Financial		3		1		4
Cultural		5		2		7
					Total	44

Category B Press: editorial

Institutional	Daily	7	Weekly	3	Total	10
Personal		7		3		10
Letters to the editor		5		2		7
					Total	27

Category C Press: reviews (theatre, books, music, dance)

	Daily	14	Weekly	3	Total	17
					Total	17

Category D Religion

Books		7
Periodicals		6
Tracts		4
	Total	17

Category E Skills and hobbies

Books		2
Periodicals		34
	Total	36

Category F Popular lore

Books	23	
Periodicals	25	
Total		48

Category G Belles lettres, biography, memoirs,etc

Books	38	
Periodicals	37	
Total		75

Category H Miscellaneous

Government documents	24	
Foundation reports	2	
Industry reports	2	
College catalogue	1	
Industry house organ	1	
Total		30

Category J Learned

Natural Sciences	12	
Medicine	5	
Mathematics	4	
Social and behavioural sciences	14	
Political science, law, education	15	
Humanities	18	
Technology and engineering	12	
Total		80

II Imaginative prose (126 samples)

Category K General Fiction

Novels	20	
Short stories	9	
Total		29

Category L Mystery and detective fiction

Novels	20	
Short stories	4	
Total		24

Category M Science fiction

Novels	3	
Short stories	3	
Total		6

Category N Adventure and western fiction

Novels	15	
Short stories	14	
Total		29

Category P Romance and love story

Novels	14	
Short stories	15	
Total		29

Category R Humour

Novels	3	
Essays, etc.	6	
Total		9
GRAND TOTAL		500

Appendix D

Contents of each major text category in the LOB Corpus

Category A	Press: reportage			
	National	daily	Political	6
			Sports	2
			Society	2
			Spot news	4
			Financial	2
			Cultural	3
	National	Sunday	Political	2
			Sports	2
			Spot news	1
			Financial	1
			Cultural	1
	Provincial	daily	Political	5
			Sports	2
			Spot news	4
			Financial	1
			Cultural	2
	Provincial	weekly	Sports	1
			Society	1
			Spot news	1
			Cultural	1
Category B	Press: editorial			
	National	daily	Institutional editorial	4
			Personal editorial	4
			Letters to the editor	3
	National	Sunday	Institutional editorial	2
			Personal editorial	2
			Letters to the editor	1
	Provincial	daily	Institutional editorial	3
			Personal editorial	3
			Letters to the editor	2
	Provincial	weekly	Institutional editorial	1
			Personal editorial	1
			Letters to the editor	1

Category C	Press: reviews		
	National	daily	6
		Sunday	5
		weekly	3
	Provincial	daily	2
		weekly	1
Category D	Religion		
	Books		9
	Periodicals and tracts		8
Category E	Skills, trades and hobbies		
	Homecraft, handiman		5
	Hobbies		5
	Music, dance		3
	Pets		1
	Sport		4
	Food, wine		2
	Travel		2
	Miscellaneous		4
	Trade, professional journals		9
	Farming		3
Category F	Popular lore		
	Popular politics, psychology, sociology		22
	Popular history		8
	Popular health, medicine		3
	'Culture'		4
	Miscellaneous		6
Category G	Belles lettres, biography, essays		
	Biography, memoirs		35
	Literary essays and criticism		6
	Arts		9
	General essays		17
Category H	Miscellaneous		
	Government documents		24
	Reports, department publications		12
	Acts, treaties		2
	Proceedings, debates		5

Other government documents		5
Foundation reports		2
Industry reports		2
University catalogue		29
Industry house organ		1

Category J Learned and scientific writings

Natural sciences		12
Medicine		5
Mathematics		4
Social, behavioral sciences		14
Psychology	4	
Sociology	5	
Demography	1	
Linguistics	4	
Political science, law, education		15
Education	4	
Politics and economics	8	
Law	3	
Humanities		18
Philosophy	4	
History	5	
Literary criticism	4	
Art	4	
Music	1	
Technology and engineering		12

Category K General fiction

Novels	20
Short stories	9

Category L Mystery and detective fiction

Novels	21
Short stories	9

Category M Science fiction

Novels	3
Short stories	3

Category N Adventure and western fiction

Novels	15
Short stories	13

Category P Romance and love story

Novels	16
Short stories	13

Category R Humour

Novels	3
Articles from periodicals	3
Articles from humorous books other than novels	3

Appendix E

Survey of the TOSCA Corpus source texts

Aldiss, B.: *Helliconia Summer*. Edition used published by Triad/Panther Books, Granada Publishing Ltd., 1985. First published by Jonathan Cape Ltd., 1983. Text classification: FSFF.

Aldiss, B.: *Seasons in Flight*. Edition used published by Triad Paperbacks Ltd., 1986. First published by Jonathan Cape Ltd., 1984. Text classification: FSTO.

Amis, K.: *Jake's Thing*. Edition used published by Hutchinson and Co. (Publishers) Ltd., 1978. Text classification: FNOV.

Ayer, A.J.: *Ludwig Wittgenstein*. Edition used published by Penguin Books Ltd., 1985. First published as *Wittgenstein* by George Weidenfeld and Nicholson Ltd., 1985. Text classification: NPHI.

Barker, C.: *Theatre Games. A New Approach to Drama Training*. Edition used published by Methuen Ltd., 1986. First published by Eyre Methuen Ltd., 1977. Text classification: NEDU.

Barker, C.: *The Damnation Game*. Edition used published by Sphere Books Ltd., 1986. First published by Weidenfeld and Nicholson Ltd., 1985. Text classification: FHOR.

Barley, N.: *The Innocent Anthropologist. Notes from a Mud Hut*. Edition used published by Penguin Books Ltd., 1986. First published by British Museum Publications Ltd., 1983. Text classification: NSOC.

Birke, L.: *Women, Feminism and Biology. The Feminist Challenge*. Edition used published by Wheatsheaf Books Ltd., Harvester Press Publishing Group, John Spiers, 1986. Text classification: NWOM.

Brazier, M.: *Medicine, Patients and the Law*. Edition used published by Penguin Books Ltd., 1987. Text classification: NMED.

Cavendish, R.: *The Tarot*. Edition used published by Chancellor Press Ltd., 1986. First published by Michael Joseph Ltd., 1975. Text classification: NMYS.

Chatwin, B.: *In Patagonia*. Edition used published by Pan Books Ltd., 1979. First published by Jonathan Cape Ltd., 1977. Text classification: NTRA.

Chatwin, B.: *On the Blackhill.* Edition used published by Pan Books Ltd., 1983. First published by Jonathan Cape Ltd., 1982. Text classification: FPSY.

Clapham, C.: *Third World Politics. An Introduction.* Edition used published by Croom Helm Ltd., 1985. Text classification: NPOL.

Clark, R.W.: *The Survival of Charles Darwin.* Edition used published by Avon Books, 1986. First published 1984. Text classification: NAUT.

Cross, R.: *Economic Theory and Policy in the UK. An Outline and Assessment of Controversies.* Edition used published by Basil Blackwell Ltd., 1985. First published by Martin Secker and Co. Ltd., 1982. Text classification: NECO.

Dawkins, R.: *The Selfish Gene.* Edition used published by Oxford University Press, 1978. First published 1976. Text classification: NBIO.

Dawkins, R.: *The Blind Watchmaker.* Edition used published by Longman Scientific and Technical, 1986. Text classification: NBIO.

Deighton, L.: *Blitzkrieg: From the Rise of Hitler to the Fall of Dunkirk.* Edition used published by Triad/Panther Books, Granada Publishing Ltd., 1985. First published by Jonathan Cape Ltd., 1979. Text classification: NHIS.

Deighton, L.: *London Match.* Edition used published by Panther Books, Granada Publishing Ltd., 1985 (special overseas edition). First published in GB by Hutchinson and Co. (Publishers) Ltd., 1985. Text classification: FTHR.

Denning, A.T.: *The Discipline of Law.* Edition used published by Butterworth and Co. (Publishers) Ltd., 1979. Text classification: NLAW.

Fowler, R.: *Linguistic Criticism.* Edition used published by Oxford University Press, 1986. Text classification: NLIN.

Fowles, J.: *The Aristos.* Edition used published by Triad Granada, 1982. First published by Jonathan Cape Ltd., 1964. Revised edition first published 1980. Text classification: NGEN.

Fowles, J.: *Daniel Martin.* Edition used published by Granada Publishing Ltd., 1985. First published by Jonathan Cape Ltd., 1977. Text classification: FNOV.

Gribbin, J.: *In Search of the Big Bang. Quantum Physics and Cosmology.* Edition used published by Bantam Books, 1986. Text classification: NPHY.

Hawkins, P.: *Introducing Phonology.* Edition used published by Hutchinson and Co. (Publishers) Ltd., 1984. Text classification: NLIN.

Herbert, J.: *The Fog*. Edition used published by New English Library, 1981. First published in an open market edition, 1975. Text classification: FHOR.

Herbert, J.: *Moon*. Edition used published by New English Library, 1985. Text classification: FSFF.

Hinde, R.A.: *Ethology. Its nature and relations with other sciences*. Edition used published by Fontana Paperbacks, 1982. Text classification: NBIO.

Hopkirk, P.: *Trespassers on the Roof of the World. The Race for Lhasa*. Edition used published by John Murray (Publishers) Ltd., 1982. Text classification: NTRA.

Horn, G.: *Memory, Imprinting, and the Brain. An inquiry into mechanisms*. Edition used published by Clarendon Press Ltd., 1985. Text classification: NPSY.

Hughes, D.: *The Pork Butcher*. Edition used published by Penguin Books Ltd., 1985. First published by Constable and Company Ltd., 1984. Text classification: FNOV.

Hughes, D.: *The Joke of the Century*. Edition published by Taplinger Publishing Co., Inc., 1986. Originally published by William Heinemann Ltd. under the title *But for Bunter*. Text classification: FHUM.

Hyland, M.: *Introduction to Theoretical Psychology*. Edition used published by The Macmillan Press Ltd., 1981. Text classification: NPSY.

Innes, M.; *Carson's Conspiracy. A Sir John Appleby Mystery*. Edition used published by Penguin Books Ltd., 1986. First published in the USA by Dodd, Mead and Company, 1984. Text classification: FCRI.

James, P.D.: *Skull Beneath the Skin*. Edition used published by Sphere Books Ltd., 1983. First published by Faber and Faber Ltd., 1982. Text classification: FCRI.

James, P.D.: *A Taste for Death*. Edition used published by Faber and Faber, 1986. Text classification: FCRI.

Jarvis, P.: *The Sociology of Adult and Continuing Education*. Edition used published by Croom Helm Ltd., 1986. First published 1985. Text classification: NEDU.

Kyle, D.: *Black Camelot*. Edition used published by Fontana Books, 1979. First published by William Collins Sons and Co. Ltd., 1978. Text classification: FTHR.

Llywelyn, M.: *Bard. The Odyssey of the Irish.* Edition used published by Sphere Books, 1985. First published by Century Publishing Co. Ltd., 1985. Text classification: FSFF.

Lodge, D.: *The Modes of Modern Writing. Metaphor, Metonymy, and the Typology of Modern Literature.* Edition used published by Edward Arnold (Publishers) Ltd., 1977. Text classification: NLIT.

Lodge, D.: *Small World. An Academic Romance.* Edition used published by Penguin Books Ltd., 1985. First published by Martin Secker and Warburg, 1984. Text classification: FHUM.

Marston, G.: *The Marginal Seabed: United Kingdom Legal Practice.* Edition used published by Clarendon Press Ltd., 1981. Text classification: NLAW.

Maskill, H.: *The Physical Basis of Organic Chemistry.* Edition used published by Oxford University Press, 1985. Text classification: NCHE.

Miles, I.: *Social Indicators for Human Development.* Edition used published by Frances Pinter (Publishers) Ltd., 1985. Text classification: NSOC.

O'Dell, D.: *Ferromagnetodynamics. The Dynamics of Magnetic Bubbles, Domains and Domain Walls.* Edition used published by The Macmillan Press Ltd., 1981. Text classification: NPHY.

Owen, D.: *A United Kingdom. An Argument and a Challenge for a Better Britain.* Edition used published by Penguin Books Ltd., 1986. Text classification: NPOL.

Pacione, M.: *Rural Geography.* Edition used published by Harper and Row Ltd., 1985. First published 1984. Text classification: NPOL.

Randle, J.: *Understanding Britain. A History of the British People and their Culture.* Edition used published by Basil Blackwell Publishers, 1981. Text classification: NHIS.

Richards, J. Radcliffe: *The Sceptical Feminist. A Philosophical Enquiry.* Edition used published by Routledge and Kegan Paul Ltd., 1980. Text classification: NWOM.

Roberts, P.: *Tender Prey.* Edition used published by Pan Books Ltd., 1985. First published by Chatto and Windus / The Hogarth Press, 1983. Text classification: FPSY.

Sacks, O.: *A Leg to Stand on.* Edition used published by Duckworth and Co. Ltd., 1984. Text classification: NMED.

Sacks, O.: *The Man who Mistook his Wife for a Hat and other Clinical Tales.* Edition used published by Simon and Schuster Inc., 1986. Text classification: FSTO.

Sharpe, A.G.: *Inorganic Chemistry.* Edition used published by Longman Group Ltd., 1986. First published, 1981. Text classification: NCHE.

Sharpe, T.: *Ancestral Vices.* Edition used published by Pan Books in association with Martin Secker and Warburg Ltd., 1983. First published by Martin Secker and Warburg Ltd., 1980. Text classification: FHUM.

Short, J.R.: *An Introduction to Political Geography.* Edition used published by Routledge and Kegan Paul Ltd., 1982. Text classification: NGEO.

Sillitoe, A.: *The Storyteller.* Edition used published by W.H. Allen and Co. Ltd., 1979. Text classification: FNOV.

Sillitoe, A.: *The Second Chance and Other Stories.* Edition used published by Jonathan Cape Ltd., 1981. Text classification: FSTO.

Silverman, P.: *Animal Behaviour in the Laboratory. Behavioural tests and their interpretation illustrated mainly by psychopharmacology in the rat.* Edition used published by Chapman and Hall Ltd., 1978. Text classification: NBIO.

Stevens, R.T.: *Flight from Bucharest.* Edition used published by Fontana Books, 1978. First published by Souvenir Press Ltd., 1977. Text classification: FROM.

Swann, D.: *Competition and Consumer Protection.* Edition used published by Penguin Books Ltd., 1979. Text classification: NECO.

Tayler, R.J.: *Galaxies: Structure and Evolution.* Edition used published by Wykeham Publishers Ltd., 1978. Text classification: NPHY.

Thomas, C.: *The Bear's Tears.* Edition used published by Sphere Books Ltd., 1985. First published by Michael Joseph Ltd., 1985. Text classification: FTHR.

Walker, E.: *A Summer Frost.* Edition used published by Grafton Books (Collins), 1986. First published 1985. Text classification: FROM.

Weldon, F.: *Praxis.* Edition used published by Hodder and Stoughton, 1984. First published 1978. Text classification: FNOV.

Weldon, F.: *Polaris and Other Stories.* Edition used published by Hodder and Stoughton, 1986. First published 1985. Text classification: FSTO.

Weldon, F.: *Rebecca West.* Edition used published by Penguin Books, 1985. Text classification: NAUT.

Wharton, C.F.P.: *Problems in Cardiology.* Edition used published by MTP Press Ltd., International Medical Publishers, 1981. Text classification: NMED.

Williams, B.: *Ethics and the Limits of Philosophy.* Edition used published by Fontana Press (Collins) Ltd., 1985. Text classification: NPHI.

Wilson, A.: *The Strange Ride of Rudyard Kipling.* Edition used published by Penguin Books Ltd., 1979. First published in the USA by the Viking Press, 1978. Text classification: NAUT.

Wilson, A.: *Setting the World on Fire.* Edition used published by Martin Secker and Warburg Ltd., 1980. Text classification: FNOV.

Wilson, A.N.: *The Laird of Abbotsford. A View of Sir Walter Scott.* Edition used published by Oxford University Press, 1980. Text classification: NLIT.

Wilson, A.N.: *Wise Virgin.* Edition used published by Penguin Books Ltd., 1984. First published by Martin Secker and Warburg, 1982. Text classification: FHUM.

Wilson, A.N.: *Hilaire Belloc.* Edition used published by Penguin Books Ltd., 1986. First published by Hamish Hamilton, 1984. Text classification: NAUT.

Wilson, A.N.: *Gentlemen in England. A Vision.* Edition used published by Penguin Books Ltd., 1986. First published by Hamish Hamilton, 1985. Text classification: FNOV.

Wilson, A.N.: *How Can We Know? An Essay on the Christian Religion.* Text classification: NREL.

Appendix F

Functions, categories and features, and their abbreviations

Functions

A	ADVERBIAL
ADAD	ADDITIVE ADJUNCT
ADAP	APPROXIMATING ADJUNCT
ADCO	CONNECTIVE ADJUNCT
ADNE	NEGATIVE ADJUNCT
ADRE	RESTRICTIVE ADJUNCT
AJHD	ADJECTIVAL HEAD
AJPO	ADJECTIVAL POSTMODIFIER
AJPR	ADJECTIVAL PREMODIFIER
APP	APPOSITIVE
AVB	AUXILIARY VERB
AVHD	ADVERBIAL HEAD
AVPO	ADVERBIAL POSTMODIFIER
AVPR	ADVERBIAL PREMODIFIER
CF	FOCUS COMPLEMENT
CJ	CONJOIN
CLOP	CLEFT OPERATOR
CM	COMMUNICATED MESSAGE
CO	OBJECT COMPLEMENT
COOR	COORDINATOR
CS	SUBJECT COMPLEMENT
CVB	CLEFT VERB
DI	DISCOURSE ITEM
DSRP	DISCONTINUOUS REPORT
DT	DETERMINER
DTCE	CENTRAL DETERMINER
DTDE	DEFERRED DETERMINER
DTPE	PREDETERMINER
DTPO	DETERMINER PHRASE POSTMODIFIER
DTPR	DETERMINER PHRASE PREMODIFIER
DTPS	POSTDETERMINER
EVB	EXISTENTIAL VERB
EXOP	EXISTENTIAL OPERATOR
EXTMA	EXTRA TEXTUAL MATERIAL
FAJPO	FLOATING ADJECTIVAL POSTMODIFIER
FAVPO	FLOATING ADVERBIAL POSTMODIFIER
FDTPO	FLOATING DETERMINER PHRASE POSTMODIFIER
FLAP	FLOATING APPOSITIVE
FNPPO	FLOATING NOUN PHRASE POSTMODIFIER
FOC	FOCUS
HDIN	HEADING INTRO
HDTL	HEADING TAIL
IMAG	IMPERATIVE AGENT
IMOP	IMPERATIVE OPERATOR
MRKUP	MARKUP
MVB	MAIN VERB
NOFU	NO FUNCTION
NOOD	NOTIONAL DIRECT OBJECT
NOSU	NOTIONAL SUBJECT
NPHD	NOUN PHRASE HEAD

NPPO	NOUN PHRASE POSTMODIFIER
NPPR	NOUN PHRASE PREMODIFIER
OD	DIRECT OBJECT
OI	INDIRECT OBJECT
OP	OPERATOR
P	PREPOSITIONAL
PART	PARTICLE
PC	PREPOSITIONAL COMPLEMENT
PMOD	PREPOSITIONAL MODIFIER
PROD	PROVISIONAL DIRECT OBJECT
PROP	PREPOSED OPERATOR
PRSU	PROVISIONAL SUBJECT
PUNC	PUNCTUATION
QTG	QUESTION TAG
RPDT	REPORTED TAIL
RPDU	REPORTED UTTERANCE
RPGI	REPORTING INSERT
RPGT	REPORTING TAIL
RPGU	REPORTING UTTERANCE
SPEC	SPECIFIER
SU	SUBJECT
SUB	SUBORDINATOR
SUBHD	SUBORDINATOR PHRASE HEAD
SUBMO	SUBORDINATOR MODIFIER
TGVB	TAGVERB
UTT	UTTERANCE
VB	VERB
VOC	VOCATIVE

Categories

ADDR	FORM OF ADDRESS
ADV	ADVERB
ADJ	ADJECTIVE
AJP	ADJECTIVE PHRASE
ART	ARTICLE
AUX	AUXILIARY
AVP	ADVERB PHRASE
CARD	CARDINAL
CL	SUBCLAUSE
CN	COMMON NOUN
COAP	APPOSITIVE CONJUNCTION
COCO	COORDINATING CONJUNCTION
CON	CONNECTIVE
CONJ	CONJUNCT
COPS	CLEFT OPERATOR STRING
COSU	SUBORDINATING CONJUNCTION
DET	DETERMINER
DTP	DETERMINER PHRASE
EOPS	EXISTENTIAL OPERATOR STRING
EXP	EXCLAMATORY PHRASE
EXT	EXTRAPOSED SENTENCE
FC	FINITE CLAUSE
FEL	FOCUSED ELEMENT
FRMEX	FORMULAIC EXPRESSION
GENM	GENITIVE MARKER
HDMO	HEADING MODIFIER

HEAD	HEADING
INT	INTERJECTION
MLV	MAIN LEXICAL VERB
MUP	MARKUP
NADJ	NOMINAL ADJECTIVE
NFC	NON-FINITE CLAUSE
NFCR	REDUCED NON-FINITE CLAUSE
NN	NOMINAL NUMERAL
NOCA	NO CATEGORY
NP	NOUN PHRASE
NPAP	APPOSITIVE NOUN PHRASE
NPG	GENITIVE NOUN PHRASE
OPP	OPERATOR PHRASE
ORD	ORDINAL
PAUX	PROCLITIC AUXILIARY
PCL	PARENTHETIC CLAUSE
PCLIT	PROCLITIC IT
PN	PRONOUN
PP	PREPOSITIONAL PHRASE
PPF	FIXED PREPOSITIONAL PHRASE
PREDP	PREDICATE PHRASE
PREP	PREPOSITION
PRIT	PROVISIONAL IT
PRN	PROPER NOUN
PRO	PROFORM
PROPP	PREPOSED OPERATOR
PRTCL	PARTICLE ITEM
PUNCM	PUNCTUATION MARK
PVB	PROCLITIC VERB
QUANT	QUANTIFIER
RESP	RESPONSIVE PHRASE
RPDS	REPORTED SEQUENCE
S	SENTENCE
SUBP	SUBORDINATOR PHRASE
TAGP	QUESTION TAG PHRASE
TXTU	TEXTUAL UNIT
VCL	VERBLESS CLAUSE
VP	VERB PHRASE

Features

ABS	ABSOLUTE
ADD	ADDITIVE
AJP	ADJECTIVE PHRASE
AP	APPROXIMATING
ASS	ASSERTIVE
AVP	ADVERB PHRASE
BLANK	BLANK
CHEAD	HEADING CLOSE
CL	CLAUSE
CLEFT	CLEFT
CO	OBJECT COMPLEMENT
COL	COLON
COM	COMMA
COMP	COMPARATIVE
COORD	COORDINATION
CXTR	COMPLEX TRANSITIVE

CQUOD	CLOSING DOUBLE QUOTE
CQUOS	CLOSING SINGLE QUOTE
CS	SUBJECT COMPLEMENT
CXDITR	COMPLEX DITRANSITIVE
CXINTENS	COMPLEX INTENSIVE
DASH	DASH
DECL	DECLARATIVE
DEM	DEMONSTRATIVE
DIA	DIALOGUE INDENTATION
DIAG	DIAGRAM
DIMOTR	DIMONOTRANSITIVE
DISCSEMI	DISCONTINUOUS SEMI
DITR	DITRANSITIVE
DO	DO
ELLIP	ELLIPSIS
ENCL	ENCLITIC
EX	EXCLUSIVE
EXCL	EXCLAMATORY
EXIST	EXISTENTIAL
EXM	EXCLAMATION MARK
EXTRA	EXTRAPOSED
FIG	FIGURE
FOR	FOR
FORM	FORMULA
GE	GENERAL
IGN	IGNORE
IGNPART	IGNORE PART
ILL	ILLUSTRATION
IMP	IMPERATIVE
IN	INTENSIFYING
INFIN	INFINITIVE
INTENS	INTENSIVE
INTER	INTERROGATIVE
INTR	INTRANSITIVE
INV	INVERTED
LET	LET
MASS	MASS
MODAL	MODAL
MOTR	MONOTRANSITIVE
NEG	NEGATIVE
NONASS	NONASSERTIVE
NP	NOUN PHRASE
OHEAD	HEADING OPEN
OD	DIRECT OBJECT
OI	INDIRECT OBJECT
ONE	ONE
OQUOD	OPENING DOUBLE QUOTE
OQUOS	OPENING SINGLE QUOTE
OTHER	OTHER
PAR	PARAGRAPH INDENTATION
PARTIC	PARTICULARIZING
PASS	PASSIVE
PAST	PAST
PASTP	PAST PARTICIPLE
PER	PERIOD
PERF	PERFECTIVE
PERS	PERSONAL
PHRAS	PHRASAL
PLU	PLURAL
POEM	POEM

POSS	POSSESSIVE
PREP	PREPOSITIONAL
PRES	PRESENT
PRESP	PRESENT PARTICIPLE
PROG	PROGRESSIVE
QM	QUESTION MARK
QUOT	QUOTATION
REC	RECIPROCAL
RED	REDUCED
REFL	REFLEXIVE
REG	REGULAR
REL	RELATIVE
SCOL	SEMICOLON
SEMI	SEMI
SING	SINGULAR
SO	SO
SONG	SONG
SU	SUBJECT
SUBJ	SUBJUNCTIVE
SUBORD	SUBORDINATE
SUBST	SUBSTANDARD
SUP	SUPERLATIVE
TABLE	TABLE
TO	TO
UNIV	UNIVERSAL
WITH	WITH
WPART	WORD PART
ZREL	ZERO RELATIVE
ZSUB	ZERO SUBORDINATE

Appendix G

An assessment of the TOSCA parser

In section 5.4.2 an evaluation of the TOSCA parser was presented. The two tables that are included in this appendix provide supplementary information to the data that were given there.

Table A gives the distribution of utterances by length for the two samples under investigation. Under F the distribution of the utterances contained in the fiction sample can be found, under N the same information is given for the non-fiction sample. In the first column the length of the utterance in number of tag units is listed. In the second column the absolute number of utterances with a particular length can be found, while in the third column the relative figures are given. The fourth and the fifth column present the cumulative numbers and figures.

In Table B an overview is given of the amount of syntactic ambiguity that the parser yielded in parsing the two samples. Under F the figures are given for the fiction sample, under N the figures for the non-fiction sample can be found. The first column lists the number of analyses that resulted from a successful parse. The largest number of analyses for a given utterance was 532. The second column lists the absolute number of utterances the analysis of which yielded a particular number of analyses. In the third column the relative figures are presented. Columns four and five list the cumulative figures, absolute and relative respectively.

ıble A: Distribution of utterances by inputlength

inputlength	# utts		% utts		cum. # utts		cum. % utts	
in # tag units	F	N	F	N	F	N	F	N
1	8	2	0.47	0.21	8	2	0.47	0.21
2	7	3	0.41	0.31	15	5	0.88	0.52
3	36	48	2.10	5.02	51	53	2.97	5.54
4	67	8	3.90	0.84	118	61	6.86	6.38
5	92	9	5.35	0.94	210	70	12.20	7.32
6	85	12	4.94	1.26	295	82	17.14	8.58
7	99	19	5.75	1.99	394	101	22.89	10.56
8	111	30	6.45	3.14	505	131	29.33	13.70
9	105	23	6.10	2.41	610	154	35.43	16.11
10	109	33	6.33	3.45	719	187	41.76	19.56
11	92	36	5.35	3.77	811	223	47.10	23.33
12	110	35	6.39	3.66	921	258	53.49	26.99
13	82	32	4.77	3.35	1003	290	58.25	30.33
14	83	25	4.82	2.62	1086	315	63.07	32.95
15	56	28	3.26	2.93	1142	343	66.32	35.88
16	71	26	4.13	2.72	1213	369	70.45	38.60
17	57	39	3.32	4.08	1270	408	73.76	42.68
18	42	30	2.44	3.14	1312	438	76.20	45.82
19	44	29	2.56	3.03	1356	467	78.75	48.85
20	53	20	3.08	2.09	1409	487	81.83	50.94
21	51	26	2.97	2.72	1460	513	84.79	53.66
22	39	27	2.27	2.82	1499	540	87.05	56.49
23	33	23	1.92	2.41	1532	563	88.97	58.89
24	19	32	1.11	3.35	1551	595	90.07	62.24
25	15	36	0.88	3.77	1566	631	90.95	66.00
26	22	23	1.28	2.41	1588	654	92.22	68.41
27	22	26	1.28	2.72	1610	680	93.50	71.13
28	14	27	0.82	2.82	1624	707	94.31	73.95
29	10	17	0.59	1.78	1634	724	94.89	75.73
30	14	28	0.82	2.93	1648	752	95.71	78.66
31	5	13	0.30	1.36	1653	765	96.00	80.02
32	8	22	0.47	2.30	1661	787	96.46	82.32
33	6	16	0.35	1.67	1667	803	96.81	84.00
34	7	15	0.41	1.57	1674	818	97.22	85.56
35	8	10	0.47	1.05	1682	828	97.68	86.61
36	4	13	0.24	1.36	1686	841	97.91	87.97
37	4	11	0.24	1.15	1690	852	98.15	89.12
38	4	9	0.24	0.94	1694	861	98.38	90.06
39	4	1	0.24	0.10	1698	862	98.61	91.32
40	4	11	0.24	1.15	1702	873	98.84	92.36
41-45	8	36	0.47	2.72	1710	909	99.31	95.08
46-50	7	15	0.41	1.57	1717	924	99.71	96.65
51-55	2	15	0.12	1.57	1719	939	99.83	98.22
56-60	1	8	0.06	0.84	1720	947	99.89	99.06
61-65	2	4	0.12	0.42	1722	951	100.00	99.48
66-70	--	1	0.00	0.10	1722	952	100.00	99.58
71-75	--	2	0.00	0.21	1722	954	100.00	99.79
76-80	--	--	0.00	0.00	1722	954	100.00	99.79
81-85	--	--	0.00	0.00	1722	954	100.00	99.79
86-90	--	1	0.00	0.10	1722	955	100.00	99.90
91-95	--	1	0.00	0.10	1722	956	100.00	100.00

Table B: Syntactic ambiguity with successful parses

# analyses	# utts		% utts		cum. # utts		cum. % utts	
	F	N	F	N	F	N	F	N
1	333	108	21.95	20.11	333	108	21.95	20.11
2	233	83	15.36	15.46	566	191	37.31	35.57
3	71	26	4.68	4.84	637	217	41.99	40.41
4	173	66	11.40	12.29	810	283	53.40	52.70
5	19	11	1.25	2.05	829	294	54.65	54.75
6	77	28	5.08	5.21	906	322	59.72	59.96
7	11	4	0.73	0.75	917	326	60.45	60.71
8	89	35	5.87	6.52	1006	361	66.32	67.23
9	20	3	1.32	0.56	1026	364	67.33	67.78
10	30	9	1.98	1.68	1056	373	69.61	69.46
11	3	6	0.20	1.12	1059	379	69.81	70.58
12	39	14	2.57	2.61	1098	393	72.38	73.18
13	3	4	0.20	0.75	1101	397	72.58	73.93
14	20	7	1.32	1.30	1121	404	73.90	75.23
15	6	2	0.40	0.37	1127	406	74.29	75.61
16	38	25	2.51	4.66	1165	431	76.80	80.26
17	2	--	0.13	0.00	1167	431	76.93	80.26
18	19	8	1.25	1.49	1186	439	78.18	81.75
19	2	1	0.13	0.19	1188	440	78.31	81.94
20	22	2	1.45	0.37	1210	442	79.76	82.31
21-30	72	25	4.75	4.66	1282	467	84.51	86.97
31-40	45	20	2.97	3.72	1327	487	87.48	90.69
41-50	29	11	1.91	2.05	1356	498	89.39	92.37
51-60	15	8	0.99	1.49	1371	506	90.38	94.23
61-70	16	5	1.06	0.93	1387	511	91.43	95.16
71-80	19	7	1.25	1.30	1406	518	92.68	96.46
81-90	12	1	0.79	0.19	1418	519	93.47	96.65
91-100	15	2	0.99	0.37	1433	521	94.46	97.02
101-125	23	6	1.52	1.12	1456	527	95.98	98.14
126-150	15	4	0.99	0.75	1471	531	96.97	98.88
151-175	10	2	0.66	0.37	1481	533	97.63	99.26
176-200	7	--	0.46	0.00	1488	533	98.09	99.26
201-250	7	3	0.46	0.56	1495	536	98.55	99.81
251-300	9	1	0.59	0.19	1504	537	99.14	100.00
301-350	4	--	0.26	0.00	1508	537	99.41	100.00
351-400	2	--	0.13	0.00	1510	537	99.54	100.00
401-450	1	--	0.07	0.00	1511	537	99.61	100.00
451-500	4	--	0.26	0.00	1515	537	99.87	100.00
>500	2	--	0.13	0.00	1517	537	100.00	100.00

Appendix H

Syntactic markers: their nature and frequency

In the syntactic pre-analysis certain constituents were marked. They are listed below, together with some examples.

Noun phrases in non-typical functions

Noun phrases that do not occur in nominal functions like subject, direct object, indirect object, subject complement, object complement, or prepositional complement, must be marked, indicating both the beginning and the end of the noun phrase. Typical examples of noun phrases that need to be marked are

NPs in adverbial position:

(1) *Last night* he got caught in the rain.
(2) She hesitated *a moment.*
(3) *A few paces from its target* the dog leapt.
(4) She dropped the pages *one by one* on to the desk.

NPs as adjectival premodifiers:

(5) *three inches* taller than Harry.
(6) *a long time* ago
(7) *a little* warmer
(8) *a mile* high

NPs as adjectival postmodifiers:

(9) worth *the risk*
(10) no nearer *a solution*

NPs as adverbial premodifiers:

(12) *face* down
(13) *a great deal* more
(14) *a little* coolly
(15) *two months* before

NPs as adverbial postmodifiers:

(16) later *that evening*

NPs as prepositional modifiers:

(17) *three quarters of an hour* into the session
(18) *a little way* behind his chair
(19) *a mile* from the city

Appositive noun phrases

The second (as well as any further) appositive in an appositive NP must be marked, indicating both the beginning and the end of the (second or further) appositive NP. For example,

(20) the word *'cause'*
(21) some flowers, *expensive hothouse blooms tied with a bow*
(22) other authors, *particularly those with psychoanalytic orientation*
(23) only one of these cases, *causality between physiological constructs*

Note that apposition may occur recursively and should be marked as such. Second or further appositives need not immediately follow an earlier appositive. For example,

(24) I wanted the truth out of him, *every last word.*
(25) Everything in it seemed insubstantial, *even the European.*

Such floating appositives are marked in the same fashion as the 'regular' appositives.

Genitive NPs

The beginning of a genitive noun phrase must be indicated. For example,

(26) *my father's* house
(27) *today's* manners

Floating determiner phrase postmodifiers

Determiner phrase postmodifiers that do not immediately follow the head they modify but are postponed are referred to as 'floating determiner phrase postmodifiers'. They must be marked, indicating both the beginning and the end of the floating determiner phrase postmodifier. For example,

(28) as few theoretical assumptions *as possible*
(29) one less behavioural consequence *than conscious mentalistic hypothetical constructs*

Adverb phrases as noun phrase premodifiers

Both the beginning and the end of adverb phrases that function as noun phrase premodifiers should be marked. For example,

(30) An *away* game.
(31) The *then* president.

Noun phrase postmodifiers

Irrespective of the category realizing the function of noun phrase postmodifier, all postmodifiers must be marked, using a marker to indicate the end of the postmodifier. By way of this marking an attempt is made to avoid the undesired ambiguity of structures that may function either as adverbial or postmodifier. For example,

(32) He saw the woman with the red skirt.

Floating noun phrase postmodifiers

Noun phrase postmodifiers that do not immediately follow the head they modify but are postponed are referred to as 'floating postmodifiers'. For example,

(33) In the next chapter another set of relations will be considered *which also depends on a separation in three kinds of hypothetical construct.*
(34) The time had come *to lay his cards on the table.*

The beginning and the end of a floating noun phrase postmodifier must be marked.

Floating adjectival postmodifiers

Adjectival postmodifiers that do not immediately follow the (adjectival) head they modify, but are postponed are referred to as 'floating adjectival postmodifiers'. For example,

(35) They have the same nature *as theoretical entities found in other disciplines.*
(36) Mentalistic hypothetical constructs have a different kind of ontological status *from that of physiological hypothetical constructs.*

The beginning and the end of a floating adjectival postmodifier must be marked.

Floating adverbial postmodifiers

Adverbial postmodifiers that do not immediately follow the (adverbial) head they modify, but are postponed are referred to as 'floating adverbial postmodifiers'. For example,

(37) In psychology, however, the concept of cause is far more complex *than in other disciplines.*
(38) The existence of unconscious mentalistic hypothetical constructs is more difficult to establish *than the presence of conscious mentalistic hypothetical constructs.*

The beginning and the end of a floating adverbial postmodifier must be marked.

Vocatives

Noun phrases that should be analyzed as vocatives receive a marker at the beginning as well as at the end of the string.

(39) *Dear Mrs C.*, you know that isn't true.
(40) "*Humphrey*," Mrs Lely said, "do tell."

Non-finite clauses

A marking is required for the beginning of those non-finite clauses that do not start off with a subordinator, or a particle followed by an overt subject. This marking serves to avoid the spurious ambiguity that would otherwise arise for multiword verbal groups, as for example in

(41) He was given the information.

Without marking two analyses are possible here, one where 'was given' is seen as one VP and 'the information' as direct object, and another where 'given the information' is taken to be an adverbial.

N.B. Particles include *for* and *with*. For example,

(42) She allowed no time *for them to comment on her last statement.*
(43) "I'd prefer to throw up *with nobody watching*, if you don't mind."

Verbless clauses

In order to restrict the amount of ambiguity that might arise from not marking verbless clauses, as well as the left-recursion we would otherwise have, all verbless clauses must be marked, indicating the beginning and the end of the verbless clause.

Parenthetic clauses

The beginning and the end of a parenthetic clause must be marked.

Table 1 gives the number of syntactic markers that were used in the syntactic pre-analysis of the two samples under discussion (see section 5.4.2). As is apparent from Table 1 the number of constituents that were marked is relatively low. Note that in 42.28% of the utterances in the fiction sample and in 19.56% of the utterances of the non-fiction sample none of the constituents was marked.

Table 1: Distribution of syntactic markers

# markers	# utts		% utts		cum. # utts		cum. % utts	
	F	N	F	N	F	N	F	N
0	728	187	42.28	19.56	728	187	42.28	19.56
1	504	199	29.27	20.82	1232	386	71.54	40.38
2	261	169	15.16	17.68	1493	555	86.70	58.05
3	119	130	6.91	13.60	1612	685	93.61	71.65
4	55	110	3.19	11.51	1667	795	96.81	83.16
5	31	63	1.80	6.59	1698	858	98.61	89.75
6	15	42	0.87	4.39	1713	900	99.48	94.14
7	6	25	0.35	2.62	1719	925	99.76	96.76
8	--	10	0.00	1.05	1719	935	99.83	97.80
9	3	8	0.17	0.84	1722	943	100.00	98.64
10	--	7	0.00	0.73	1722	950	100.00	99.37
11	--	1	0.00	0.10	1722	951	100.00	99.48
12	--	1	0.00	0.10	1722	952	100.00	99.58
13	--	--	0.00	0.00	1722	952	100.00	99.58
14	--	1	0.00	0.10	1722	953	100.00	99.69
15	--	1	0.00	0.10	1722	954	100.00	99.79
16	--	2	0.00	0.21	1722	956	100.00	100.00

Samenvatting

Dit proefschrift rapporteert over (aspecten van) het corpustaalkundig onderzoek dat gedurende de periode 1981-1989 werd verricht aan de Katholieke Universiteit te Nijmegen. In het kader van een tweetal opeenvolgende onderzoeksprojecten werden de onderzoeksinstrumenten ontwikkeld en/of aangepast voor het verrichten van corpustaalkundig onderzoek met behulp van de computer, en vervolgens toegepast bij de syntactische analyse van een corpus Engelse teksten.

In hoofdstuk 1 wordt een karakterisering gegeven van de corpustaalkunde zoals die zich gedurende de afgelopen tien jaar heeft ontwikkeld. De aandacht gaat daarbij uit naar die aspecten waarin de corpustaalkunde zich onderscheidt van andere takken van de taalkunde, alsmede de methodologische verschillen die er bestaan tussen de corpustaalkunde enerzijds en overige benaderingen van de computationele linguïstiek anderzijds. Verder wordt het Nijmeegse onderzoek zoals dat in de twee TOSCA projecten werd verricht en dat de basis vormt voor de voorliggende dissertatie geplaatst in het kader van de hierboven beschreven ontwikkelingen.

In hoofdstuk 2 wordt meer specifiek ingegaan op de rol van het corpus in corpustaalkundig onderzoek, met name het onderzoek dat zich richt op de bestudering van taalvariëteit. Factoren die een rol spelen in het onderzoek naar taalvariëteit worden besproken. Geconstateerd wordt dat veelgebruikte corpora zoals het Brown Corpus en het LOB Corpus minder geschikt zijn als basis voor onderzoek naar taalvariëteit in al zijn aspecten. Er wordt nagegaan welke de eisen zijn die aan (de samenstelling van) een corpus moeten worden gesteld wanneer men dergelijk onderzoek nastreeft. Tevens wordt uiteengezet hoe deze zijn geoperationaliseerd bij de samenstelling van het Nijmeegse TOSCA Corpus.

Hoofdstuk 3 beschrijft de uitgangspunten die bepalend zijn voor het opstellen van een formele grammatica en beargumenteert de keuze voor het gehanteerde formalisme, dat van Extended Affix Grammar (EAG). Voorts wordt een globale beschrijving gegeven van de structuur van de grammatica zoals die voor het Engels werd ontwikkeld.

Aan de hand van de beschrijving van coordinatie en gapping enerzijds en de noun phrase anderzijds wordt in hoofdstuk 4 geïllustreerd op welke wijze het EAG formalisme kan worden aangewend bij het formaliseren van taalkundige beschrijvingen zoals die

voorhanden zijn in grammaticale handboeken van het hedendaags Engels.

In hoofdstuk 5 tenslotte, wordt ingegaan op de problemen die zich hebben voorgedaan bij het schrijven van de grammatica en de toepassing ervan in het daaropvolgende analyseproces. Voorts wordt een bespreking gewijd aan de wijze waarop een analysesysteem zoals beschreven moet worden geëvalueerd, en wordt deze evaluatie toegepast op het analysesysteem waarover in dit proefschrift wordt gerapporteerd.

Curriculum Vitae

Nelleke Oostdijk was born in Nijmegen on 24 September 1958. In 1977 she received her Atheneum A diploma from Dominicus College, Nijmegen and began her studies in English Language and Literature at the University of Nijmegen, specializing in Modern English Linguistics. She graduated (cum laude) from this university in 1983. After graduation she was appointed as a research assistant to the first TOSCA project, which she had joined earlier as a student assistant. Upon conclusion of this project in 1985 she was employed (part-time) by the Dutch Research Council, while she remained employed (also part-time) by Nijmegen University as a temporary staff member. She has been employed on a full-time basis by Nijmegen University since 1989. As a lecturer in the Department of Language and Speech she continues to be involved in research in the field of corpus linguistics and in the teaching of a number of courses in this area.